Automatic Extraction of Rarely Explored Materials and Methods Sections from Research Journals using Machine Learning Techniques

Kavitha Jayaram[1], Prakash G[2]
Department of Computer Science and Engineering
Amrita School of Engineering
Amrita Vishwa Vidyapeetham, Bengaluru, India

Jayaram V[3]
Solid States and Structural Chemistry Unit
Indian Institute of Science, IISc
Bengaluru, India

***Abstract*—The scientific community is expanding by leaps and bounds every day owing to pioneering and path breaking scientific literature published in journals around the globe. Viewing as well as retrieving this data is a challenging task in today's fast paced world. The essence and importance of scientific research papers for the expert lies in their experimental and theoretical results along with the sanctioned research projects from the organizations. Since scant work has been done in this direction, the alternative option is to explore text mining by machine learning techniques. Myriad journals are available on material research which throws light on a gamut of materials, synthesis methods, and characterization methods used to study properties of the materials. Application of materials has many diversified areas, hence selected papers from "Journal of Material Science" where "Materials and Methods" sections contains names of the method, characterization techniques (instrumental methods), algorithms, images, etc. used in research work. The "Acknowledgment" section conveys information about authors' proximity, collaborations with organizations that are again not explored for the citation network. In the present articulated work, our attempt is to derive a means to automatically extract methods or terminologies used in characterization techniques, author, organization data from "Materials and Methods" and "Acknowledgment" sections, using machine learning techniques. Another goal of this research is to provide a data set for characterization terms, classification and an extended version of the existing citation network for material research. The complete dataset will help new researchers to select research work, find new domains and techniques to solve advanced scientific research problems.**

***Keywords*—Data-mining; rule-based; machine-learning; term extraction; classification; materials and methods; acknowledgment**

I. Introduction

Citation networks have been well analyzed both syntactically as well as structurally but there is a strong need for semantic analysis for these networks. Citation analysis as the most significant area of bibliometric that has been studied using the Page Ranking algorithm for a long time and there has been a great deal of research work in this direction [1]. Citation sentiment analysis is used to determine sentiment polarity of clinical trial papers using n-gram and sentiment lexicon features on annotated corpus [2]. A summary of a corpus of research papers, domain-independent structural relations between abstracts and domain of scholarly medical articles, state-of-the-art deep learning baseline was constructed and has been reported [3]. Given a particular paper of interest, CiteSeer can display the context of how the paper is cited or indexed in subsequent publications with a summary of the paper in electronic format [4]. The semantic analysis of paper abstracts is a good start for annotating papers using Natural Language Processing (NLP) with semantic metadata and for increasing the general representation and visualization of the key concepts within a given domain [5]. Here they discuss and analyze the text mining techniques and their applications in diverse fields [6]. The collaboration of productive authors based on the topics, collaborative effort, highly cited articles, etc. would identify the relationship between two specific nodes that can reveal scholarly communication patterns (i.e., collaboration or knowledge diffusion, copyright transfer) with finer granularity [7]. The author proposes a mathematical model that matches empirical acknowledgment data closely for citation patterns which give cognitive interdependence among disciplines [8]. A function of appreciation using the acknowledgment section within academia of the instrumental and normative significance has been presented [9]. A content-based image retrieval system that extracts, image features from journal papers using a supervised learning algorithm has been explained in [10]. An overview of the principles and methods of automatic term recognition of significant elements have been presented [11]. From conference proceedings and journal papers information extracted like dataset, content, and basis of extraction summarized in Appendix I. Scientometrics researchers use structural/syntactic information from a bibliographic network for qualitative analysis of the same network [12]. Some common aspects like the dataset used, methods, the most focused problem in a particular field, frequently used algorithms, hot areas such as analysis of research trends have been extracted [13]. Fig. 1 represents the existing citation network focuses on citation count and co-authorship hence it mainly contains four nodes namely Venue, Paper, Term, and Author/co-authors [12].

The "Abstract" section contains the best ratio of keywords per total of words, which contains research, material, methods used and challenges faced, but many times they do not include methods. Hence, the next most important findings of the research are expressed in "Materials and Methods" section such as experimental techniques, instrumental methods, algorithm, figures, etc. Acknowledgments section is used to express

appreciation between researchers, direct or indirect collaborators, and the contribution of external people or organizations. These aspects of the citation network are important and are needed to be explored to improve author proximity, affiliations, and funding organizations that contribute to academic or industrial research. It is found that the automatic identification of methods and acknowledgment influences the citation network. The modified citation network where few other nodes like "organization" have been included to study the author collaboration, method and dataset nodes from Materials and Methods section to analyze compounds or materials is shown in Fig. 2. It is clear that extracting the above mentioned important information from the "Journal of Material Science" and incorporating it into the existing citation network give us new ways to look at the authors' communities, collaborators from organizations and institutions in material research. This paper describes the automatic extraction of materials, characterization techniques, instrument-related terminologies, acknowledged by authors (organization) using Machine Learning (ML) techniques.

This paper is organized as follows: Section 2 covers the implementation of the algorithm, tools, framework, and work executed in the present research. Section 3 explains the experiments, results, and discussion. Section 4 summarizes the present work and future research which can be laid upon the work. In the last Section, we acknowledge the research collaborators. Appendix I include a list of reference papers where the information is extracted from the present published research papers. Appendix II gives lists of sample research papers from the research journals with title, materials, and characterization techniques. Appendix III gives a list of characterization methods used to investigate the results presented in the materials and methods section of the journal publications.

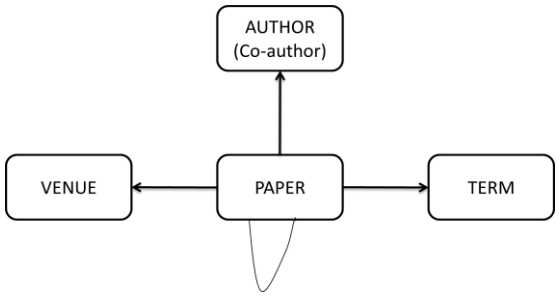


Fig. 1. Current Citation Network Analysis.

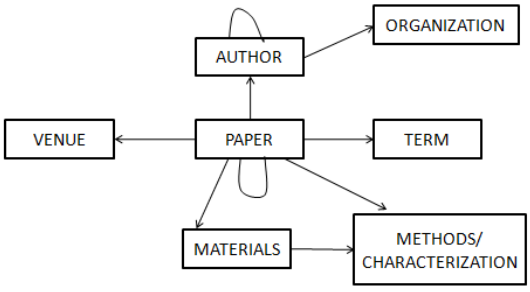


Fig. 2. Modified Citation Network for Materials and Methods from Scientific Journals.

II. Implementation

The building of a heterogeneous network includes different types of entities and incorporating into the currently existing citation network. The main work comprises finding out the entity mentioned (characterization techniques and organization names) from the research work. The implementation details mainly focuses on the "Materials and Methods" section which includes materials, methods, figures, micrographs, images, and material characterization obtained from different instrumental methods. To analyze "Material and Methods" section, it is required to convert the extracted information into a format compatible with usual heterogeneous citation networks. New node types such as "Algorithm, technique, method", "characterization, measurements, instruments", "Images or figures" and “Organization funding" form the key semantic components of the research work. To the best of our knowledge, no such work has been done with "Materials and Methods" and "Acknowledgment" sections from material research journals using machine learning techniques.

The statistical approach generally uses information such as term frequency, term-document frequency, inverse term-document frequency, etc. for extracting the important entities and mentioned phrases. Named entity recognition (NER) is also a problem that attempts to find out mention, author, organization, place, etc. Although their extraction is very good, it is limited to particular classes and does not have any model to mention terminologies and acronym. Although a lot of work has been done on domain-specific term extraction and named entity extraction for particular classes, the method keywords extraction has not been explored. Both rule-based and ML approaches to find methods were mentioned in a scientific research document to extract important techniques and methods used in biomedical research [14].

Scientific documents are mostly available in PDF format, which is semi-structured and not tagged, unlike HTML, also 'text' in them is usually arranged in multiple rows and columns. Many tools are available to extract text from PDFs, but when documents come with multiple rows and columns like tables, figures, etc., text extraction tool is not good enough[11]. A rule-based approach that is leveraged to extract the required sections is proposed using regular expressions and was reported in [15]. Single-word does not represent an entity, but a sequence of words does, support vector machine (SVM), linguistic-based techniques for entity extraction generally uses part-of-speech (POS) tagging and the dependencies of the words upon each other [16]. The ML algorithms used are Naïve Bayes' classifier, decision tree, and maximum entropy classifier. Extraction of a vast number of terminologies and acronyms from the "Materials and Methods" section is not an easy task. New methods and techniques are being used and named with new emerging problems. Using PDFBox and TET tools, the extraction of spatial co-ordinates and formatting information of text has been completed. In the present research work, automatic extraction of entities like text, single nouns and compound nouns has been carried our using a machine learning approach instead of linguistic methods.

[1] http://www.pdflib.com/products/tet/

Primarily the 'text' has been extracted using PDFBox text extraction tool, and then the co-ordinates of words and lines in the documents were calculated. This helps in calculating the coordinates of the line where the section name is extracted using regular expression, starting from one section to the next section, using a regular expression. Further "Materials and Methods" and "Acknowledgment" sections were also extracted individually from PDF into a text format using a section extraction algorithm as presented in Fig. 3. After extracting the required sections the terminologies like names of the materials, characterization techniques**,** methods, authors, organization, etc. are extracted from the text file. Terminology and acronym for materials and characterization techniques from the sample journals are listed in Appendix II.

```
Open PDF document
While lines: start to end do
if line_text == secName then
   set co-ordinates and page; break
end
end
while lines : sec_pagedo
if line_coord>= sec_coordthen
extract lines
end
if (line_txt == secName) and  (line_coord>=
sec_coordOR
line_page>sec_page) then
  break;
end
end
```

Fig. 3. Section Extraction Algorithm [12].

The following two categories were considered for extracting acronym:

Category 1: Methods ending with keywords (such as analysis or scope) eg.: Energy Dispersive X-Ray Analysis/spectroscopy (EDX or EDS).

Category 2: Methods do not have any keywords. eg.: X-ray Diffraction (XRD).

New dataset was created by selecting data from nearly 800 research papers, where it contains methods and characterization techniques. The method mentions in the dataset representing the characterization techniques (or Instrumental methods), algorithm, theory, model are considered in the form of nouns. In category 1, methods are extracted using regular expressions to create the training dataset. Hence POS (Part of Speech) was used for tagging to extract names of all methods, though they fall into any of these two categories. When data falls into category 2, supervised machine learning algorithms have been used, for which a good quantity of training dataset is required. All the relevant characterization techniques and abbreviations are listed in Appendix III. Different classification algorithms were run over datasets and evaluated by precision, recall, and F1-score techniques using the following formulas [17]:

$$\boldsymbol{precison} = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

$$\boldsymbol{recall} = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

$$\boldsymbol{Precision} = \frac{t_p}{t_p + f_p}, \qquad \boldsymbol{Recall} = \frac{t_p}{t_p + f_n}$$

$$\boldsymbol{F} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$\textbf{Balance accuracy} = \frac{\text{TPR} + \text{TNR}}{2}$$

$$\textbf{TPR} = \frac{\text{tp}}{\text{tp} + \text{fn}} \text{ and } \textbf{TNR} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

Whereas for classification, the following terms are used to compare the results of the classifier: the term t_p is true positives, t_n is true negatives, f_p is false positives, f_n is false negative, further, TPR is term positive rate and TNR is term negative rate. Precision is the fraction of relevant instances among the retrieved instances, while recall fraction of the total amount of relevant instances that were actually retrieved. F1-score is the harmonic mean of precision and recall. TPR and TNR are statistical classification for a confusion matrix or error matrix. The terms positive and negative refer to the classifier's prediction (expectation) and true and false terms refer to the prediction corresponds to the external judgment (observation) [18].

III. Results and Experiments

Present experiments were performed on system configuration having 128 GB RAM by using Python 3.0 with nltk and also Java as the programming language.

Data pre-processing was performed before collecting training data, such as removing all stop words, commas, semicolons, newlines (which were unnecessarily present because the data was extracted from pdfs). The papers were downloaded from an official website of "Journal of Material Science (JMS)". The text contained in the documents was extracted using PDFBox tool. Even this tool is not found to be very promising in retaining the structure of the extracted text. Since the data is in PDF format, it is a difficult task to use all the information available in the research documents. Therefore, data pre-processing becomes an important and time consuming task. Spatial coordinates of the words to form the lines and to keep the lines in correct order are also an important task. After working on many methods, good results were achieved by a supervised classification method approach. Summary of noun phrases from journal papers and Wikipedia entries term sequence are listed in Appendix III.

Both "Materials and Methods" and "Acknowledgment" sections are derived using the regular expression based rules. Previous work on section extractions shows that regular expressions achieve 100% precision and 67% recall for extracting Acknowledgment section [19]. The proposed analysis shows that same approach works for the Materials and

Methods section too. Hence regular expressions and spatial co-ordinates are used to extract both sections of the research paper. NLP technique is used to extract sentences having the materials name, algorithm methods or characterization, measurement, etc. words. StandfordCoreNLP tool is used for named entity recognition [20]. Using these entities a list of the most widely used methods or simulation work done in material research are listed in Appendix III. Named Entity Recognition (NER) is used to extract sentences from different papers to find out methods, characterization, algorithms, people / authors, and organizations.

Noun phrases available in the research paper are searched from the Wikipedia entries. The summaries of the term sequences were collected while rejecting the sequences that were not available in Wikipedia entry. Along with these entries documents were clustered into five classes using Linear Discriminant Analysis (LDA), the list of methods predicted in a paper (w.r.t. materials used for research) is shown in Fig. 4(a). Using nearest neighbor and supervised methods the corresponding classes were assigned to dominant topics. With the important extracted information, the Citation Network is extended to provide dataset related to collaborators and authors due to newly introduced nodes in the network. The results obtained also include a new dataset for characterization techniques from the research paper.

The main goal is to extract the characterization or methods from the research papers. Initially, about 100 term sequences from various research papers were manually tagged as methods (characterization methods). These 100 terms were extracted from research papers and Wikipedia entries terms using rule-based regular expressions techniques based on machine learning and NLP methodologies. Subsequently all the relevant stop words, commas, semicolons, newlines (which are unnecessarily present because of the data extracted from pdf's) were removed from the extracted text. Although many of the problems that arose owing to the pdf's extraction were addressed, few problems remain unsolved. Few problems like, unnecessary spacing between few words, some non-ASCII characters, and distortion of table data are attributed as the primary reasons for messing up in text data. These mistakes could have a detrimental effect on the output. The features extracted from the text are automatically run by the program where the positive and negative class term sequences are encountered.

The process of searching term sequences in the whole document set for creating the training data manually while considering positive and negative term sequences with a ratio of 5:3, but the ratio obtained was about 1:9. This is the class imbalance problem that occurred due to the specificity of the positive term sequences and all the general noun phrases coming into negative class. This class imbalance problem was resolved by applying entity clustering for sampling negative class instances, where the ratio was about 6:4. Fig. 4(a) shows the list of materials classified as carbon, graphite, silica, electronic, and high-temperature materials (HTC) along with a sub-classified list of few compounds selected from the Journals. Fig. 4(b) shows the list of characterization techniques from different instrumental methods like XRD, SEM, TEM, etc. including few simulation and ML done on the selected materials. Summary of the characterization methods used to analyze the material selected from the "Materials and Methods" section of the Journal are listed in Table I. All these classifications are considered while solving problems using machine learning techniques.

The materials and characterization techniques extracted from the Journal gives the following conclusion. The bar graph in Fig. 5(a) shows that TiO_2, Graphene materials appear in more research papers compared to other compounds. However, in Fig. 5(b) shows XRD and SEM instruments used extensively as characterization techniques for material analysis. Our results from the machine learning techniques reveals the statistics of materials not explored by the researchers and the type of methods not used for characterization of materials including simulation work. The present research work provides a dataset for materials and methods for selecting particular area of research by the scientific community.

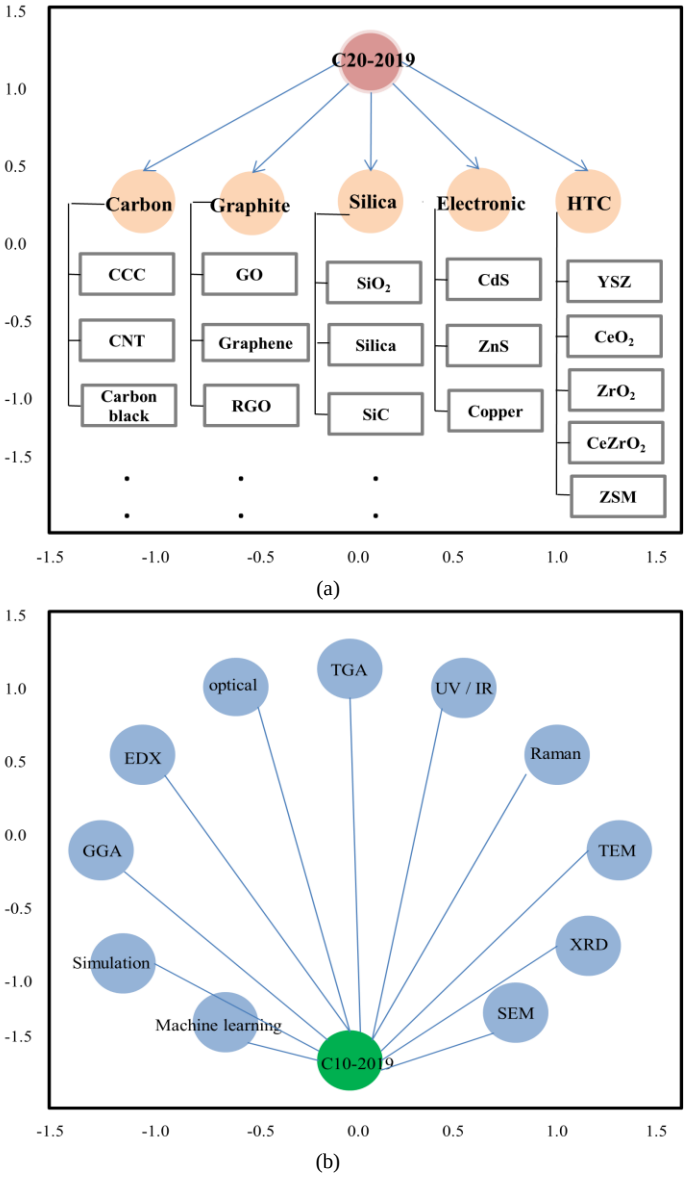
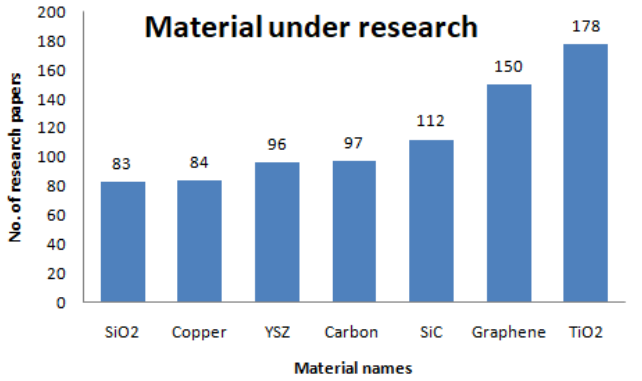


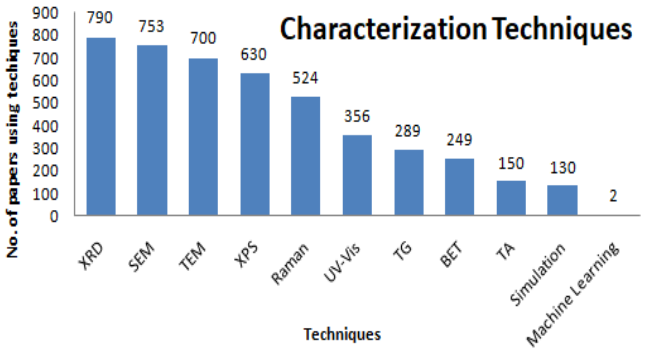
Fig. 4. Dataset Selected for the Classification from the Journal of Material Science (a) Materials (b) Methods.

TABLE I. LIST OF SELECTED MATERIALS ALONG WITH INSTRUMENTAL METHODS USED FOR CHARACTERIZATIONS FROM THE JOURNAL PAPERS

Sl no.	Material	Characterization techniques
1	Graphene	TEM, HRTEM, EELS, ITC, XPS, ES, Raman, UV-Vis, DSC, TGA, Vickers hardness indenter, Zeta potential, Material Studio (MD simulation)
2	SiO_2	SEM, TEM, Raman, XRD, XPS, Raman, Simulation
3	TiO_2	TEM, SEM, TIFR, Raman, XRD, XPS
4	YSZ	CALPHAD, TA, SEM, EDX, XRD, XPS,
5	Carbon black	XRD, SEM, TEM, Raman, IR camera, UV-Vis-NIR
6	Copper	Raman, SEM, UV-Vis, FESEM, XRD, BET, TEM, EIS, MD Algorithm
7	Organic Polymer	NMR spectra, FTIR, TGA, XPS, XRD, SEM, BET, UV-Vis, Fluorescence spectra
8	SiC	SEM, Raman, SRIM simulation
9	Nanofibers	SEM, XRD, TEM, BET, TGA, Raman
10	Ceramics	Gibbs energy, DSC, TGA, XRD, SEM, RADIANT



(a)

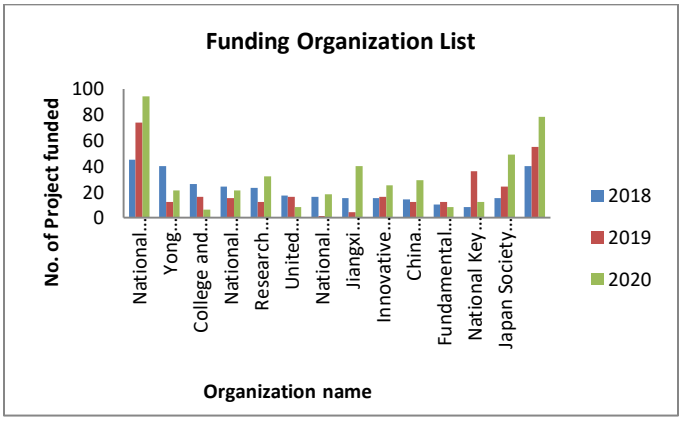


(b)

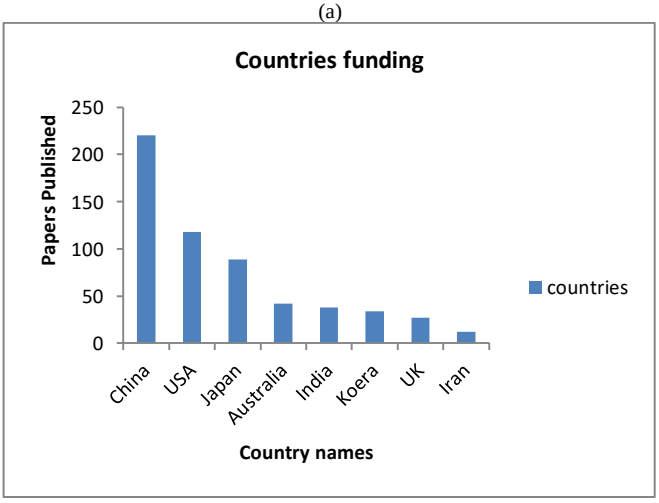
Fig. 5. Data Extracted from Journal of Material Science Papers (a) Top 7 Materials, (b) Top 10 Characterization Techniques.

The clustering is a frequency, which contains names, short names, and abbreviations using different alternatives for some well-known organization in acknowledgment section are added in the training dataset. Top 14 organizations which are acknowledged in Material Research Journal were analyzed for research publication. An analysis of the acknowledgment section along with organization and country names extracted from the Journal of Material Science for the past three years is shown in Fig. 6(a). Few selected funding agency are listed in Table II. According to the graph, NNSFC (National Natural Science Foundations of China) is the most acknowledged Chinese organization, involved in funding the most research projects. The analysis shows China published most research papers followed by the USA and other countries as shown in Fig. 6(b).

In summary, the results show that China published more research papers, and NNSFC funded the maximum project in past three years. The comparison shows the number of the research paper published by different countries in the past three years. This period can be extended to more number of years to validate our machine learning techniques. Once the author and organization parameters are extracted, built a social network of the acknowledgment section, and the snapshot of the social network is shown in Fig. 7.



(a)



(b)

Fig. 6. Projects Funded from different Countries are Selected from Acknowledgment Section, (a) Bar Graph of different Organization Funded the Project from Past 3 Years Vs Number of Projects, (b) Bar Graph of Project Funding Countries Vs Number of Research Paper Published.

TABLE II. LIST OF FUNDING AGENCY

Sl no.	Funding agency
1	National Natural Science Foundation
2	Yong Teachers Scientific Research
3	College and University Key Project
4	National Funds for Distinguished Young Scientists
5	Research Fund of University
6	United Innovation
7	National Research Agency
8	Jiangxi Scientific / Education Fund
9	Innovative Research Team in University
10	China Scholarship Council
11	Fundamental Research Funds for Central University
12	National Key Research and Development Program
13	Japan Society for the Promotion of Science
14	National Science and Technology cooperation Funds

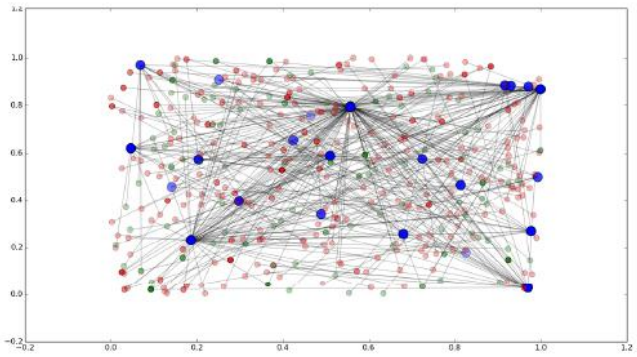


Fig. 7. Social Network of Acknowledgment Section Processed for 100 Published Research Papers. Blue = Papers, Red = Organization, Green = Person.

Novelty and evaluation is a very important part of research work. Precision and recall are two extremely important model for evaluation metrics. While precision refers to the relevant percentage of results, recall refers to the percentage of total relevant results correctly classified by the algorithm. F-1 score is the harmonic mean of precision and recall. Both precision and recall are important to solve problems; one can select a model that maximizes the F1 score. To check the correctness of the predicted method terms, 20 research documents from the same journal were selected at random. Manually extracted methods and characterization techniques were used from the "Materials and Methods" sections and the results were compared with classification algorithms [21, 22]. Precision, Recall and F1-scores are different classification algorithms used to predict characterization techniques and organization names. Also we have computed dataset using LDA, NBS and LIBLINEAR to evaluate classification algorithms. LDA (Latent Dirichlet Allocation) is a generative probabilistic model for collections of discrete data such as text corpora [23]. NBC (Neighborhood Based Clustering) discovers clusters based on the neighborhood characteristics of data [24]. LIBLINEAR (Library for Large Linear Classification) uses logistic regression and linear support vector machines which is very efficient on large sparse datasets [25]. The values listed in Table III are the measured dataset from the classification algorithms; which concludes **LDA**, NBC and LIBLINEAR (SVM) and gives better F1 scores. Overall results show the novelty of our research work which generates and establishes tagged training dataset extracted from the materials and methods section using ML technique, which supports researchers to select advanced research topics.

TABLE III. EVALUATED DATASET MATRIX OF THE CLASSIFICATION ALGORITHMS

Method list	Recall	Precision	F1-Score
NBC	0.68	0.32	0.41
LIBLINEAR	0.68	0.30	0.41
Decision Tree	0.83	0.24	0.37
MaxEnt	0.72	0.26	0.38
Random	0.48	0.16	0.24
LDA	**0.54**	**0.34**	**0.41**

IV. CONCLUSION

Our analysis shows there are plenty of hidden information in each section of research journal papers. The extracted information can be used to extend the currently existing Citation Network. "Materials and Methods" and "Acknowledgments" are the least explored aspects of Scientometrics of Material Science Research papers. The methods and characterization from the "Materials and Methods" section, people and organizations acknowledged from the "Acknowledgments" section were extracted from "Journal of Material Science" and revealed important insight. A new researcher or a beginner can get an idea of material as well as the characterization methods used for completion of the research work. They can also understand which material, techniques are least explored for new research domains to proceed. Gives adequate information about the ongoing research problems, researchers are interested to find out the country, collaborators, and to propose new joint research project form different funding agencies. Future work involves extracting the "Abstract" and "Results" sections from scientific research journals. These two sections helps in summarizing the classification of completed research work, figures can be classified according to image quality and instrumental methods used for characterization of the materials. The complete dataset will help new researchers to select research work, find new domains and techniques to solve advanced scientific research problems.

ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors. I express my heartfelt gratitude to Prof. M Narasimha Murty and Mr. Rohit Kumar from Computer Science & Automation department, IISc, Bangalore, India, for their constructive criticisms and timely directions which led to the successful completion of this work.

REFERENCES

[1] Elliot J Yates and Louise C Dixon, "PageRank as a method to rank biomedical literature by importance", Source Code of Biology and Medicine, vol. 10, pp. 16-25, 2015.

[2] Jun Xu, Yaoyun Zhang, Yonghui Wu, Jinggi Wang M S, Xiao Dong M D and Hua Xu, "Citation sentiment analysis in clinical trial papers", AMIA Annual Symposium Proceedings Archive, pp. 1334-1341, November 2015.

[3] Arthur Brack, Jennifer D Souza, Anett Hoppe, Soren Auer and Ralph Ewerth, "Domain-independent extraction of scientific concepts from research articles", Springer Nature Switzerland AG, pp. 251-266, 2020.

[4] C Lee Giles, Kurt D Ballacker and Steve Lawrence, "Citeseer: An automatic citation indexing system", Proceedings of the third ACM conference on Digital libraries, New York, pp. 89-98, 1998.

[5] Ionut Cristian Paraschiv, Mihai Dascalu, Stefan Trausan-Matu, Philippe Dessus, "Analyzing the semantic relatedness of paper abstract", 20th International Conference on Control System and Science, pp. 759-764, 2015 [IEEE].

[6] Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha and Fakeeha Fatima, "Text mining: techniques, applications and issues", International Journal of Advanced Computer Science and Applications,Vol. 7, No. 11, pp. 414-418, 2016.

[7] M E J Newman, "Coauthorship networks and patterns of scientific collaboration, colloquium", The National Academy of Sciences of the USA, Vol. 101, pp. 5200-5205, 2004 [PNAS].

[8] Charles H Davis and Blaise Cronin, "Brief communication acknowledgments and intellectual indebtedness: A bibliometric conjecture", Journal of the American Society for Information Science, vol. 44, no..10, pp. 590-592, 1993.

[9] Blaise Cronin, "Acknowlegdement trends in the research literature of informantion science", Journal of Documentation, 2001, Vol. 57, No. 3, pp. 427-433, 2001.

[10] B Akshaya, S S Sruthi Sri, A Niranjana Sathish, K Shobika, R Karthika and Latha Parameswaran, "Content-based image retrieval using hybrid feature extraction techniques", Springer Nature Switzerland AG, pp. 583-593, 2019 [ISMAC-CVB].

[11] Bin Umino, "Methods of automatic term recognition – A review", Proceedings of COLING, Mumbai, pp. 1211-1222, 2012 [Technical Papers].

[12] Madian Khabsa, Pucktada Treeratpituk, and C Lee Giles, "Ackseer: a repository and search engine for automatically extracted acknowledgements from digital libraries", Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, USA, pp. 185-194, 2012.

[13] Hospice Houngbo and Robert E Mercer, "Method mention extraction from scientific research papers", Proceedings of COLING, pp. 1211-1222, 2012.

[14] Bei Yu, "Automated Citation Sentiment Analysis: What can we learn from biomedical researchers", ASIS&T Proceedings, vol. 50, no. 1, pp. 1-9, 2013.

[15] Xiaoyu Tang, Qingtian Zeng, Tingting Cui and Zeze Wu, "Regular expression-based reference metadata extraction from the Web", 2nd symposium on web society, pp. 346-350, 2010 [IEEE].

[16] Didier Bourigault, "Surface grammatical analysis for the extraction of terminological noun phrases", Proceedings of the COLING, pp. 977-981, August 1992 [NANTES].

[17] David M W Powers, "Evaluation: From precision, recall and F-factor to ROC", Informedness, Markedness & Correlation, Technical report SIE-07-001, pp. 1-24, 2007.

[18] C Lee Giles and Isaac G Councill, "Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing", The National Academy of Sciences of the USA, vol.101, no. 51, pp. 17599-17604, 2004.

[19] Christopher D Mining, Mihai Surdeanu, Sohn Baner, Jenny Finkel, Steven J Bethard and David McClosky, "The Standford corenlp natural language processing toolkit", Proceedings of 52nd annual Meeting of Association for Computational Linguistics, USA, pp. 55-60, June 2014.

[20] Kamal Sarkar, Mita Nasipuri and Suranjan Ghose, "Machine learning based keyphrase extraction: comparing decision trees, Navie Bayes and artifical neural networks", J Inf Process Syst, India, vol. 8, no. 4, pp. 693-712, December 2012.

[21] Kavitha Jayaram, Sangeeta , "A review: Information extraction techniques from research papers", ICIMIA, India, pp. 56-59, February 2017 [IEEE].

[22] David M Blei, Andrew Y Ng and Michael Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research,vol. 3, pp. 993-1022, 2003.

[23] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin, "LIBLINEAR: a library for large linear classfication", Journal of Machine Learning Research, 2008, 9: 1871-1874.

[24] Shuigeng Zhou, Yue Zhao, Jihong Guan and Joshua Huang, "A neighborhood-based clustering algorithm", Springer-Verlag Berlin Heidelberg, pp. 361-371, 2005 [LNAI 3518].

APPENDIX I. LIST OF EXTRACTED INFORMATION FROM SELECTED JOURNAL PAPERS

Sl no	Paper title	Dataset	Extracted Content	Basis of Extraction
1	IE from Biomedical Literature: Methodology, Evaluation and an Application (2003)	Biomedical dataset	Biological terms in doc, identify the common concepts in group of genes	Dictionary, clustering of genes,
2	Automatic extraction of titles from general documents using machine learning (ACM 2005)	Internet of Microsoft, DotGov, DotCom	Titles	Improving of search ranking results in doc retrieval by extracted titles
3	Mining knowledge from text using information extraction (2005)	Book amazon.com	Abstract knowledge, concrete data	Knowledge, patterns
4	Extracting procedures from text (2007)	Public recipe web site	Procedures and graphs	Building large knowledge
5	Automatic extraction and processing of document references (2007)	CRF-based approach	References , names of related documents	Search these doc in sys DB, create links to respective documents
6	Opinion holder extraction from author and authority viewpoints (2007)	MPQA corpus	Named entity extraction	Opinion of author and authority viewpoint
7	Entity categorization over large documents collections (ACM KDD, 2008)	Web-data	Entities (people, movies, painter, writer)	Categorizing extracted entities
8	Corpus study of kidney-related experimental data in scientific papers (2009)	Quantitative kidney Database (QKDB)	Experimental data	Automate extraction of experiment and result section
9	Automatic creation of a technical trend map from research papers and patents (ACM 2010)	NTCIR-8 patent mining task (Japanese data)	Technology (algo, tools, materials, data)	Creating Technical trend map (recall and precision)

10	Link Analysis in mind maps: a new approach to determining doc relatedness (2010)	Maps	Reference	Doc recommender if doc A and B are refereed by mind map then related
11	Contextual Information Extraction in Research Articles: A case of developing contextual RDF data for ESWC papers (ACM 2011)	European Semantic Web Conference (ESWC)	Paragraphs containing citations and classified sentences	Automatic context identification. Author work, cited work used by various authors, searching citation sentences, export data in different formats
12	Citation analysis and keyword mining based on full text extraction of scientific literature (ACM 2012)	ACM digital library	Citation, bibliometric analysis (domain context graph)	Domain knowledge graph and analyzing interrelation of publications for research direction
13	A comparison of metadata extraction techniques for crowdsourced bibliographic metadata management (2012)	e-prints, Mendeley dataset	Authors	Conditional Random Fields and SVM
14	Machine Learning based keyphrase extraction: comparing decision trees, naïve bayes, and artificial neural network (2012)	Downloaded from websites of journals Springer, Elsevier, (economics, law, medical)	Comparing different keyphrase extraction	Comparison of methods
15	ARTIC: metadata extraction from scientific papers using a two-layer CRF model (2014)	100 scientific papers from IEEE, Elsevier, Springer and ACM	Title, author names, emails, affiliations and venue info	Conditional Random Fields to extract metadata from scientific papers
16	Recommendation of newly published research papers using belief propagation (2014)	DBLP dataset	Citation information	Recommend most interesting newly published paper
17	Scientific monitoring by mining scientific papers (2014)	PDF and HTML	Semantic annotation	Relation between organizations or topics
18	Automatic extraction of main thesis documents fields using decision trees (2015 International conference on Computational Science and Computational Intelligence)	Downloaded thesis, 65 documents (converted to word format)	Title, abstract, authors	Facilitate solving the research problem process. Structuring thesis document to help research to access knowledge easily
19	Competing Algorithm Detection from Research Papers (ACM 2016)	DBLP (small dataset)	Algorithm names (name entity extraction)	Competing algorithm (ranking based on number of comparisons
20	Insights from mining eleven years of scholarly paper publications in requirements engineering (RE) series of conferences (2016)	551 papers from Requirements Engineering (RE)	Topics frequently co-occurring and connected terms, co-author	University-industry collaborations, internal and external collaborations
21	PDFFigures 2.0: mining figures from research papers (2016) (more fig more citations)	Introduce a new dataset of comp science papers	Figures and tables, captions	Component analysis, online search engine, correlates figure citation.
22	Extracting code segments and their descriptions from research articles (2017 IEEE/ACM international conference on mining s/w repositories (MSR)	IEEE digital libraries	Code example	Functionality and properties

APPENDIX II. LIST OF MATERIALS AND CHARACTERIZATION TECHNIQUES SELECTED FROM "MATERIALS AND METHODS" SECTION FROM THE SAMPLE JOURNALS

Sl. No.	**Title**	**Material**	**Characterization techniques**
1	Evolution of phase during heating of metastable beta titanium alloy Ti-15Mo	Ti	TEM
2	Mesoscale simulations of shockwave energy dissipation via chemical reaction	Polymer	MD Simulation (ChemDID), shock tube
3	Oxygen ion mobility and conductivity prediction in cubic yttria-stabilized zirconia single crystals	Yttria-stabilized zirconia (YSZ)	CALPHAD
4	Nitrogen-doped porous carbon derived from imidazole-functionalized polyhedral oligomeric silsesquioxane	3-Chioropropyltrimethoxysilane	NMR, FTIR, TEM, SEM, XRD, XPS and Raman spectrometer
5	Ultrahigh strength in nanocrystalline materials under shock loading	Copper	Shockwave, simulation, TEM
6	Facile preparation of Mg-doped graphitic carbon nitride composites as a solid base catalyst for Knoevenagel condensations	Mg-doped g-C_3N_4	XRD, SEM, TEM, XPS
7	Rational design of CuO nanostructures grown on carbon fiber fabrics with enhanced electrochemical performance for flexible supercapacitor	CuO	FESEM, XRD, BET, EIS
8	Magnetoresistance of graphite intercalated with cobalt	Pyrolytic graphite	XRD
9	Non-catalytic behavior of SiO_2 fine powders in presence of strong shockwaves for aerospace application	SiO_2	Material Shock Tube (MST), XRD, SEM, TEM, HRTEM, XPS

10	Thickness dependence of photoresponsive properties at $SrTiO_3$-based oxide heterointerfaces under different strains	$SrTiO_3$-based oxide	Atomic force microscopy (AFM), X-ray reflectivity (XRR)
11	Response of microstructure to annealing in-situ Cu-Nb microcomposite	Nb with Cu-Nb microcomposite	SEM, TEM
12	Poly (ε-caprolactone)/cellulose nanocrystal nanocomposite mechanical reinforcement and morphology: the role of nanocrystal pre-dispersion	Cellulose nanocrystal (CNC)	Young's modulus, TEM
13	Transparent heat insulation coatings with high selective shielding ability designed with novel superstructures of copper sulfide nanoplates	CuS superstructures	SEM, EDS, FESEM, XRD, FTIR, UV-Vis
14	The effect of applied voltage on the corrosion resistance of MgO-C refractories	MgO-C	Composition using EDS from SEM facility
15	Experimental investigation on response and failure modes of 2D and 3D woven composites under low velocity impact	Woven composites	Olympus stereo microscope for damage impact, C-scan
16	Aging response on the stress corrosion cracking behavior of wrought precipitation-hardened magnesium alloy	Magnesium alloy	TEM
17	Flexible tuning of hole-based localized surface plasmon resonance in roxbyite Cu1.8S nanodisks via particle size, carrier density and plasmon coupling	Nanocrystals (NCs)	TEM, SEM, UV-Vis, XRD
18	Ab-initio calculations of $CaZrO_3$ (011) surfaces: systematic trends in polar (011) surface calculations of ABO_3 perovskites	Polar $CaZrO_3$	Ab-initio calculation (Simulation)
19	Tensile testing of aged flexible unidirectional composite laminates for body armor	Composite laminates	SEM
20	On the binary Sb-Sn system: ab-initio calculation and thermodynamic remodeling	Alloy of Sb-Sn	Ab-initio calculation VESTA software

APPENDIX III. LIST OF ABBREVIATIONS, METHODS AND MEASUREMENT TECHNIQUES

Sl. no.	Abbreviations	Methods	Measurements
1	AFM / SFM	Atomic Force Microscopy / Scanning Force Microscopy	Very-high resolution type of scanning probe microscopy, force measurement, topographic imaging and manipulation
2	BET	Brunauer Emmett Teller	The theory aims to explain the physical adsorption of gas molecules on a solid surface and to measure porosity and surface specific area of nano materials.
3	CALPHAD	Theoretical method	A CALPHAD thermodynamic database allows the calculation of the equilibrium state of "real" engineering materials.
4	CT	Computed tomography	It enables a three-dimensional representation of the internal and external structure of objects with a detailed detect-ability which goes down into the micrometer range.
5	DTA	Differential thermal analysis	The material under study in an inert atmosphere is made to undergo identical thermal cycles while recording any temperature difference between sample and reference.
6	DFT (PBE-DFT)	Density Function Theory (Perdew-Burke-Ernzerh of DFT)	Computational quantum mechanical modeling method used in physics, chemistry, and materials science to investigate the electronic structure (or nuclear structure) (principally the ground state) of many-body systems, in particular atoms, molecules, and the condensed phases.
7	EA	Electrochemical analyzer	It provides trace metal analysis, trace organic analysis, computer-controlled cyclic voltammeter, and chronoamperometry techniques.
8	EDX or EDS	Energy Dispersive X-Ray (EDX) Energy Dispersive Spectroscopy (EDS)	Chemical microanalysis technique used for elemental analysis in conjunction with SEM.
9	EELS	Electron Energy Loss Spectroscopy	Material is exposed to a beam of electrons with a known kinetic energies. Some of the electrons will undergo inelastic scattering, which means that they lose energy and provides information on unoccupied energy level.
10	EIS	Electrochemical Impedance Spectroscopy	Study of doped spinal manganese cathode oxide materials synthesized for Li-ion batteries.
11	ELS Zeta potentiometer	Electrophoretic Light Scattering	In contrast, streaming potential measurements, no movement of the liquid is generated, but the movement of the particles is used to measure suspended particle size in fluids.
12	EPR / ESR	Electron Paramagnetic Resonance / Electron Spin Resonance	Spectroscopic technique that detects species that have unpaired electrons, a surprisingly large number of materials have unpaired electrons.
13	ES / OES	Emission Spectrometer / Optical Emission Spectroscopy	A rapid method for determining the elemental composition of a variety of metals and alloys. Chemical analysis labs are equipped to evaluate the properties of the material.

14	FESEM / SEM	Field Emission Scanning Electron Microscope	An advanced microscope offering increased magnification and the ability to observe very fine features at a lower voltage than the SEM.
15	FTIR	Fourier Transform Infrared Spectroscopy	An analytical technique used to identify organic (and in some cases inorganic) materials. The technique is used to obtain an infrared spectrum of absorption and emission spectra of solid, liquid, and gas.
16	Gibbs free energy	Calculated	Calculates the Thermodynamic potential of the material.
17	HRTEM /TEM	High Resolution Transmission Electron Microscopy/ Transmission Electron Microscopy	High-resolution TEM offers resolution down to the Angstrom level and gives information on the atomic packing, rather than just the morphology. Particle growth can also be studied using TEM.
18	Hybrid rheometer		Accurate measure of frequencies, material types, and experimental designs.
19	Laser diffraction particle size		Light scattering method for particle size analysis of covering a wide range from submicron to millimeter scale.
20	MST / ST	Material Shock Tube/Shock Tube	It is a device consisting of driver and driven sections separated from a metal diaphragm, used to accelerate the test gas in supersonic and hypersonic speed, upon stopping it produces high temperature and pressure used to interact with materials at the end of the shock tube
21	NMR	Nuclear magnetic resonance (NMR) spectroscopy	Used to determine the structure of organic molecules in solution and study molecular physics, crystals as well as noncrystalline materials. Also used in advanced medical imaging techniques, such as magnetic resonance imaging (MRI).
22	Optical parameter oscillator /OM	Optical microscope	The basic optical microscope, improves resolution, uses visible light, easy to develop.
23	RADIANT	RADIANT ferroelectric testing	Characterizing non-linear materials. Precision and accuracy have been the driving force behind the engineering of test equipment and thin ferroelectric film components.
24	Raman Spectra	Raman Spectroscopy	Commonly used in chemistry to provide a structural fingerprint by which molecules can be identified. The technique typically used to determine vibrational modes of molecules, although rotational and other low-frequency modes of systems may also be observed.
25	RT-MS	Room Temperature-Monochromator Spectrometer	A monochromator produces a beam of light with a very narrow bandwidth of light of single color. It is widely used for spectroscopic analysis of sample materials. The incident light from the light source can be transmitted, absorbed, or reflected through the sample.
26	SPS	Syndiotactic Polystyren	SPS techniques are refractory metals and intermetallics, oxide, and non-oxide ceramics. The particles constituting the powders before consolidation tend to decrease their surface energy by desorption of chemical species, once introduced inside the SPS chamber.
27	TCSPC	Time-correlated single-photon counting	Fluorescence lifetimes, occurring as emissive decays from singlet-state, approximated in time region from picoseconds to nanoseconds.
28	TF Analyzer	Thin-film analyzer	The most sophisticated analyzer of electro-ceramic materials and devices. The test equipment is based on a modular idea, where four different probe heads can be connected to the same basic unit. Each of the four-probe heads offers different characterization methods.
29	TG	Thermogravimetric Analysis	Thermal analysis in which the mass of a sample is measured over time as the temperature changes.
30	USAXS	Ultra-small-angle X-ray Scattering Spectrometer	SAXS and USAXS belong to a family of small angle X-ray scattering techniques that are used in the characterization of materials. This instrument can record data at smaller angle, to resolve and probe larger dimension objects.
31	UV-Vis	Ultra Violet – Visible Spectroscopy	It is absorption spectroscopy, measurement of attenuation of a beam of light after it passes through a sample or after reflection from the sample surface.
32	VSM	Value-Stream Mapping	Analyzes flow of materials.
33	XPS / ESCA	X-ray photoelectron spectroscopy or Electron Spectroscopy for Chemical Analysis	Widely used for surface analysis technique because it can be applied to a broad range of materials and provides valuable quantitative and chemical state information from the surface of the material being studied.
34	XRD	X-ray Powder Diffraction	The analytical technique primarily to identify crystal structure, unit cell, particle size and strain measurement.
35	XRF	X-ray Fluorescence Spectrometer	A non-destructive analytical technique used to determine the elemental composition of materials. XRF analyzers determine the chemistry of a sample by measuring the fluorescent (or secondary) X-ray emitted from a sample when it is excited by a primary X-ray source.
36	XRR	X-ray reflectivity	It is a analytical technique using reflected beam of x-rays from flat surface, measured for the intensity of x-rays reflected in direction to understand surface-sensitivities