Unsupervised Technique for Automatically Extracting Components of References

Kalpana Uppada, M. Kranthi Kiran, Sridhar Seepana, S. Jahnavi, K. Gipson Nikil

Abstract: ***The automatic extraction of bibliographic data remains a difficult task to the present day, when it's realized that the scientific publications are not in a standard format and every publications has its own template. There are many "regular expression" techniques and "supervised machine learning" techniques for extracting the entire details of the references mentioned within the bibliographic section. But there's no much difference within the percentage of their success. Our idea is to seek out whether unsupervised machine learning techniques can help us in increasing the share of success. This paper presents a technique for segregating and automatically extracting the individual components of references like Authors, Title of the references, publications details, etc., using "Unsupervised technique", "Named-Entity recognition"(NER) technique and link these references to their corresponding full text article with the assistance of google.***

Key Words: Regular Expression technique, supervised machine learning, Bibliography, References, Unsupervised technique, Name-Entity recognition.

I. INTRODUCTION

Researchers usually download various research papers while performing ground work and analysis. To search an article or research paper, we will either manually type or copy the title of the paper and retrieve the research paper with the help of search engines like google. While performing literature survey, most of the analysts prefer to use snowball sampling technique in which the researchers check for the most appropriate references in the primary article and get the corresponding full text article using google. As researchers will have numerous articles downloaded in their local file system, it is laborious to check all the files manually. Instead, they prefer to download the reasearch paper again. By this method our main motto of reference linking is not concentrated as we are promoting redundancy. There are many reference management softwares such as Mendely, Zotero, SodhanaRef, etc to extract the metadata such as title of the paper, author names, publication details, etc inorder to perform the search and download the required research paper but still taking the title part of each and every reference and searching it manually in the reference management software is also a difficult task for researchers.

Revised Manuscript Received on April 27, 2020.

Kalpana Uppada , Bachelors, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam.

Dr. Mandava Kranthi Kiran , Assistant Professor, GITAM Institute of Technology, GITAM Deemed to be University.

Sridhar Seepana , Bachelors, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam

Jahnavi Setti , Bachelors, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam

Gipson Nikil , Bachelors, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam

Toeliminate this problem and to reduce the manual strain for the researcher we have "An Approach Towards Establishing Reference Linking in Desktop Reference Manager" [1] which helps in automation of reference article linkage using strict regular expression technique for which the performance is 78.44% and "SodhanaRef" [2] which has reference article linking along with providing a semantic search for which the performance is 67%. We also have "A Strategy for Automatically Extracting References from PDF Documents" [4] which used supervised machine learning with annotations and regular expression technique. The increase in performance is 74% which is very less. So to improve the performance, we are demonstrating an unsupervised approach for automatically extracting the components of the references and thereby linking the references to a search engine.

A. Machine Learning Machine

Learning is the ability to boost the behavior of the machine based on previous experiments [6] or experiences and to learn from data with respect to some class of tasks and performance measures by choosing training data and how to represent the target function by choosing an appropriate algorithm to infer a target function.

The concept of Machine Learning [7] is categorized into two types:

1.Supervised Machine Learning 2.Unsupervised Machine Learning [9]

B. Unsupervised Machine Learning

In Unsupervised machine learning, the machine is neither trained nor the data is labelled. We will not train the machine with particular patterns but will let the machine to learn on its own[24]. Unsupervised machine learning is further categorized into two types of algorithms:

Clustering 2.Association

Clustering: A Cluster means a group of data. Clustering is a process of grouping the data with similarities [12]. Data in the same group or cluster are more similar while compared to other clusters.

Association: Association [12] is the process of finding the relationship between data variables in large datasets.

C. Natural Language Processing

Natural language processing (NLP) [10] gives machine the ability to read, understand, analyze and derive conclusions from human languages. It mainly deals with how computers process and analyze natural languages[11]. Using semantics with natural language processing helps us to derive the exact meaning of the text or sentence by processing word by word and then finding the relation between the words to produce accurate and exact results. Applications of NLP involve speech recognition, text classification, sentimental analysis, etc.

II. PROPOSED ARCHITE,CTURE

A research paper in pdf form is considered as an input and the whole text from the pdf is copied into a text file. From the text file, only the references are identified and are extracted into a new text file. Each individual reference is spotted from the text file and is segregated to give the author name, title and proceedings as output. Using NLP, author names and proceedings are eliminated in order to produce a list of titles as output. As a result of the above process, a list of titles are extracted as output into a new text file. The researcher can easily refer to a particular paper using the title instead of searching all references and can give the desired title to a search engine like google which obtains the corresponding full text article as output.

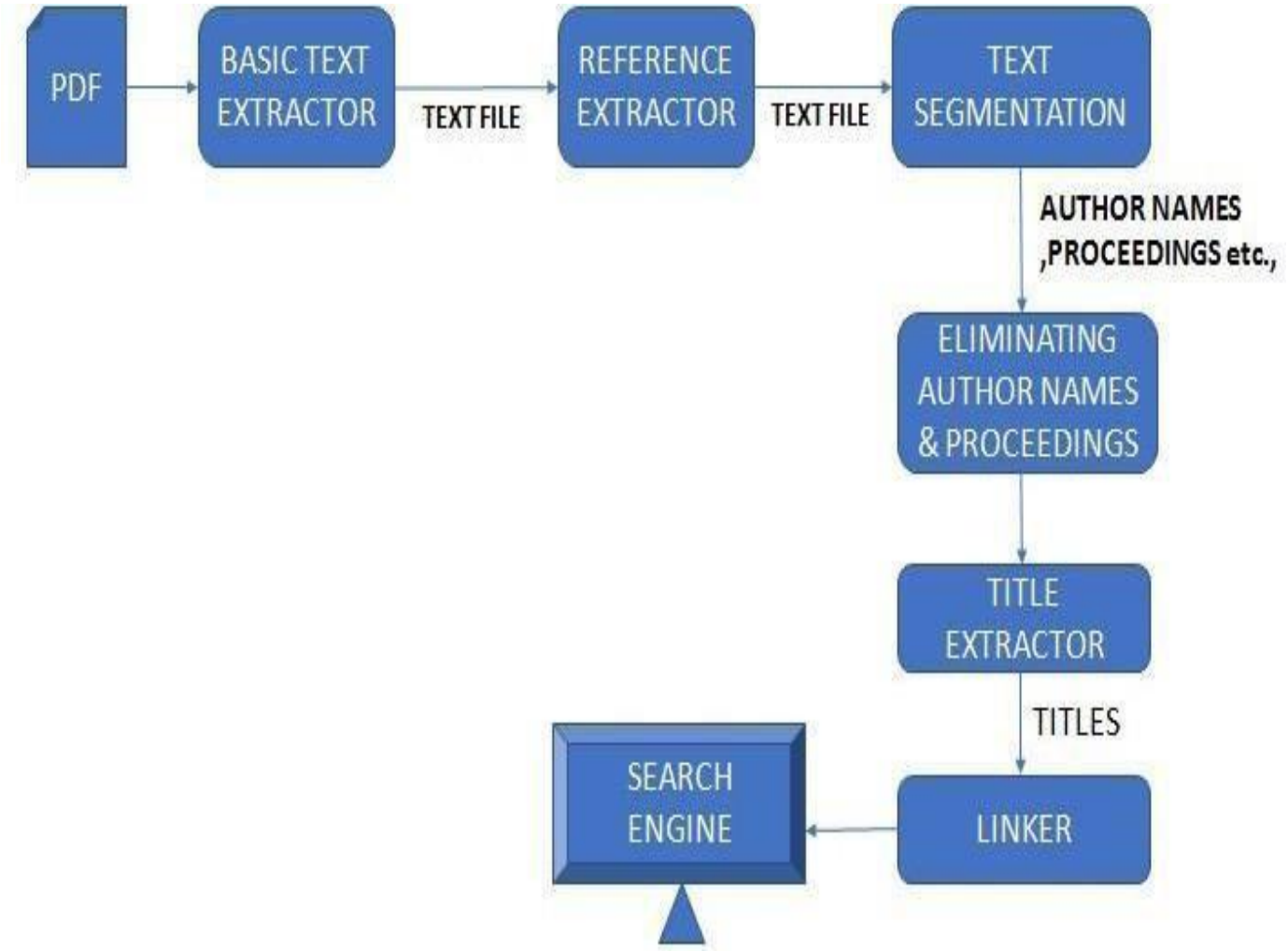

Fig.1: Architecture of proposed system

III. BASIC TEXT EXTRACTION

A. Identifying References

Initially the full text PDF file of a research paper is taken as input in order to identify the keyword “References” or “REFERENCES” using regular expressions. This is done by matching the keyword with every word in pdf. Once the match is found, its respective page number is identified and is used for further implementation.

B. Extraction of References into a text file

Only the page number where the match for the word ‘Reference’ is found is taken as input and now match for the word reference is searched in that particular page. The text which proceeds the word ‘Reference’ is ignored and text succeeding the word reference along with the word reference is extracted into a new text file. Nowthis new text file will only contain the references of the research paper in PDF form which is given as input.

IV. AUTHOR IDENTIFICATION

Various methods have been implemented as a part of our research. At first, we thought of identifying authors with the help of parts of speech tagging using natural language processing and then apply K-means clustering algorithm and to divide the input into three clusters of authors, titles and proceedings but as many words are not proper nouns and some of the proper nouns which are part of our titles are recognized as proper nouns, and hence there is high chance of misclassifying words in clusters. So applying K-means will not yield us good results. Secondly, we thought of using google scholar. Google scholar provides us the facility to search and view the author's profile. As almost all authors would be listed in google scholar, we have searched all NNP’s in the authors profile page of google

scholar. We have found that all the author names are identified correctly, but we have encountered a problem when we have words like computer (which is a noun) while searched in the author's profile, yielded wrong results (if the author is of computer science department, keyword match for the computer is done in the profile). So even this method did not give us better results.

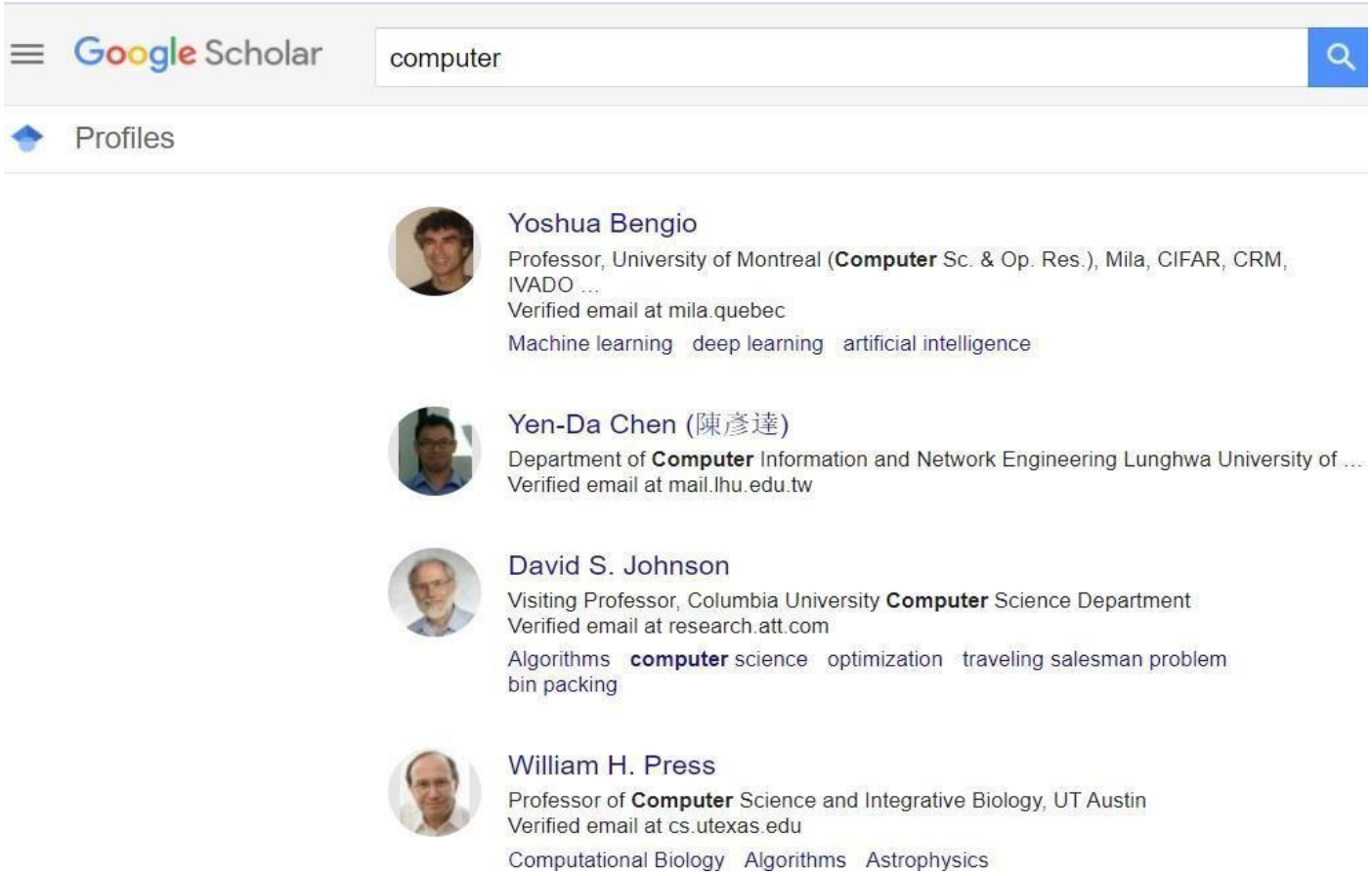

Fig.2: Results of google scholar author search

So, finally, we proceeded with NER technique. The text file which only contains references is taken as input. Identifying the authors is a three step process which can be done using natural language processing (NLP). Initially, we removed all the punctuations using and then we tokenized the whole text of references to give parts of speech tagging. At last NNP's are identified and are removed.

A. Removing punctuations

As we have intended to follow the unsupervised process, we are initially eliminating punctuations. By using punctuations, we can easily identify and segregate individual components of references. Titles can be identified by using double quotes (" "). Text preceding the titles i.e. text before first dot (.) and double quotes ("") will mostly be author name and text succeeding titles will be proceedings but using these regular expressions comes under supervised technique. So to purely proceed with our key goal of identifying titles using unsupervised process, Punctuations such as ! ( ) - [ ] { } ; . : '" ' \ , <> /
? @ # $ % ^ & * _ ~ are identified and removed.

B. Text tokenization and POS tagging

Tokenization is the process of dividing a stream of data into words and each segmented individual word is known as token. POS tagging [13] which is part of named-entity recognition is given to each token to identify its parts of speech. The output of POS tagging
[14] would be the word along with its parts of speech. For example consider the word 'and'. Now 'and' would be an individual token and parts of speech tagging will be given to 'and'. So the result will be (and,CC) where
CC represents conjunction.

C. Identifying NNP

NNP represents only the proper nouns. The words with parts of speech (POS) tagging NNP are identified and are discarded. Therefore, by extracting the NNP's we would be able to recognize all the author names from the references as author names will be proper nouns.

V. TITLE EXTRACTION AND LINKING

After eliminating all the author names, we are left with the title and proceedings. To extract titles, we have to discard proceedings. Extracting key words such as proceedings, international, conference, journal are recognized and text succeeding these keywords are removed. Finally, titles are extracted. Now, the research paper that the researcher intends to look at can be downloaded or viewed by giving it to a search engine like google.

VI. EVALUATION AND RESULTS

Accuracy: The ratio of count of exact outputs to the total number of inputs. It is also defined as ratio of an error to the range of possible output (Full-scale output) values.

Average Accuracy = Number of exact titles extracted/ Total number of titles

We have analyzed different formats of research papers such as IEEE, ACM, ELSEVIER, SPRINGER and other papers too. The results are as follows:

Table-I: Accuracy of titles extracted from various references of various research papers

FORMAT		Total no. of References	Total no. of Correctly Extracted References	ACCURACY
IEEE	PAPER -1	9	6	66.6
	PAPER -2	11	7	63.6
ACM	PAPER-1	9	7	77.7
ELSEVIER	PAPER-1	19	14	73.6
SPRINGER	PAPER -1	5	5	100
	PAPER -2	10	7	70
OTHER JOURNALS	PAPER -1	29	20	68.9
	PAPER -2	32	24	75

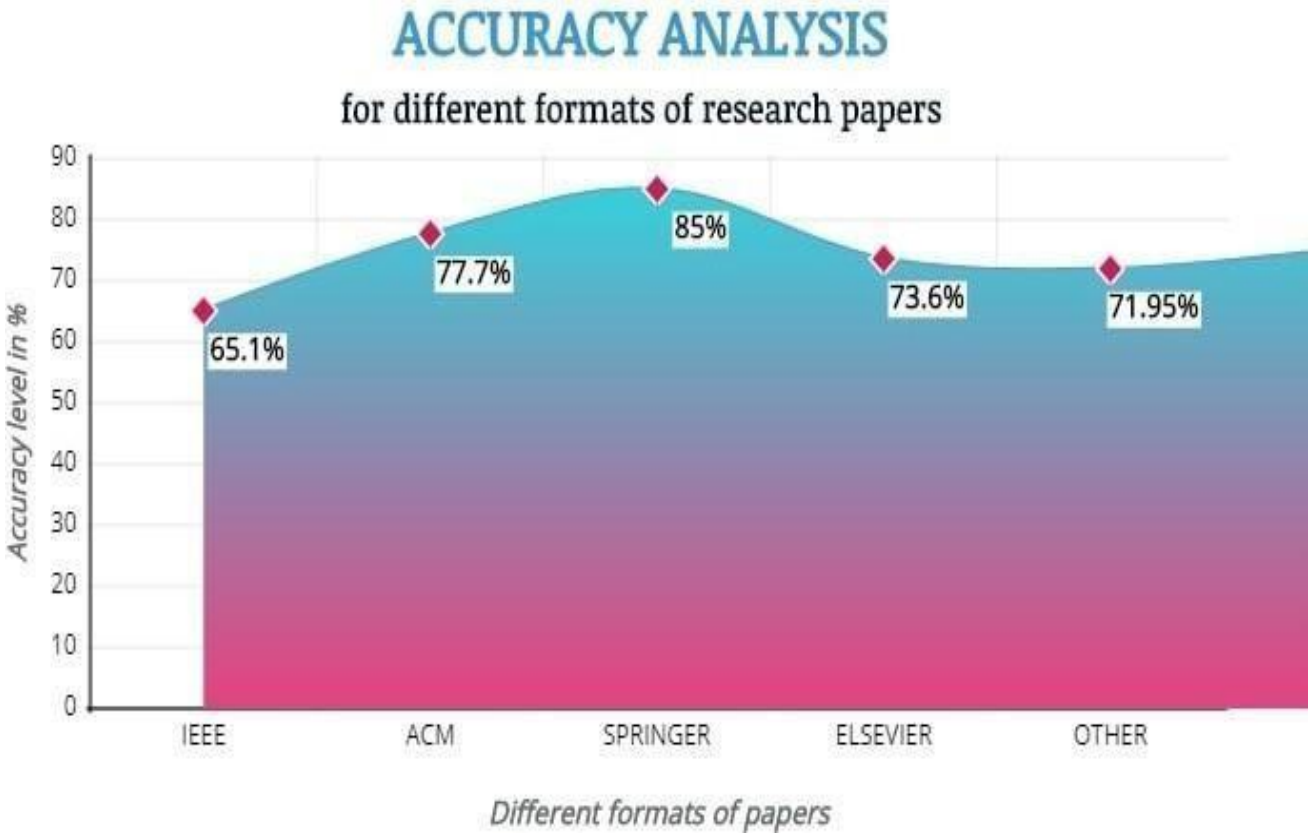

Fig.3: Accuracy analysis in graph format

So, analyzing different papers, we have attained an average accuracy of 72.91% in extracting titles from the components of references.

VII. CONCLUSION AND FUTURE WORK

Researching with unsupervised technique for automatically extracting the titles from the references, we have encountered that the accuracy is 72.91%.We could improve the percentage of SodhanaRef which is our inspiration. However if the machine is trained i.e., with the usage of supervised technique "A Strategy for Automatically Extracting References from PDF Documents" the accuracy in extracting titles is 74% which is very close to ours. Since the machine is not trained with any regular expressions and due to misclassification of proper nouns, there is no much improvement in the performance compared with the supervised technique. We have observed that the research papers which contain less number of references resulted in high accuracy. Though individual accuracy of papers like SPRINGER, ACM and ELSEVIER are high but as we analyze the average accuracy for several formats of papers we attained 72.91%. Further, we believe that our goal can be achieved if there is a proper way for identifying named nouns(only person names) instead of identifying all the proper nouns in Named-Entity recognition. Use of regular expression along with POS tagging (which comes under semi supervised technique) might also increase the performance which we want to research further.

REFERENCES

1. Mandava Kranthi Kiran, K. Thammi Reddy. "An Approach Towards Establishing Reference Linking in Desktop Reference Manager". Journal of Information & Knowledge Management,Vol. 17, No. 3 (2018).
2. Mandava Kranthi Kiran, K. Thammi Reddy."SodhanaRef: a reference management software built using hybrid semantic measure". International Journal of Engineering & Technology, 7 (2) (2018) 495505.
3. Matta chenet."Identify and extract entities from bibliography references in a free text",2017.
4. Neide Ferreira Alves, Rafael Dueire Lins, Maria Lencastre." A Strategy for Automatically Extracting References from PDF Documents". 10th IAPR International Workshop on Document Analysis Systems,2012.
5. P.D Turney, M.L Litman."Unsupervised Learning of Semantic Orientation from a Hundred-Billion-WordCorpus",2002.
6. F. Musumeci et al., "An Overview on Application of Machine Learning Techniques in Optical Networks," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383-1408, Secondquarter 2019.
7. P. Godefroid, H. Peleg and R. Singh, "Learn&Fuzz: Machine learning for input fuzzing," 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE), Urbana, IL, 2017, pp. 50-59.
8. M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 24, no. 3, March 2002.
9. A. S. Park and J. R. Glass, "Unsupervised Pattern Discovery in Speech," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 186-197, Jan. 2008.
10. M. Ibrahim and R. Ahmad, "Class Diagram Extraction from Textual Requirements Using Natural Language Processing (NLP) Techniques," 2010 Second International Conference on Computer Research and Development, Kuala Lumpur, 2010, pp. 200-204.
11. C. Arora, M. Sabetzadeh, A. Goknil, L. C. Briand and F. Zimmer, "Change impact analysis for Natural Language requirements: An NLP approach," 2015 IEEE 23rd International Requirements Engineering Conference (RE), Ottawa, ON, 2015.
12. Hsiangchu Lai and Tzyy-Ching Yang, "A group-based inference approach to customized marketing on the Web integrating clustering and association rules techniques," Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Maui, HI, USA, 2000.
13. M. Nghiem, D. Dinh and M. Nguyen, "Improving Vietnamese POS tagging by integrating a rich feature set and Support Vector Machines," 2008 IEEE International Conference on Research, Innovation and Vision for the Future in Computing and Communication Technologies, Ho Chi Minh City, 2008, pp. 128- 133.
14. A. Belaid, L. Pienon and N. Valverde, "Part-of-speech tagging for table of contents recognition," Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, Barcelona, Spain, 2000, pp. 451-454 vol.4.
15. Dominika Tkaczyk, Pawe? Szostek, Piotr Jan Dendek, Mateusz Fedoryszak and ukasz Bolikowski,"CERMINE — automatic extraction of metadata and references from scientific literature", April 2014 with 1,155 Reads DOI: 10.1109/DAS.2014.63,Conference: 11th IAPR International Workshop on Document Analysis Systems.
16. .R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," 2010 IEEE Symposium on Security and Privacy, Berkeley/Oakland, CA, 2010, pp. 305-316.
17. J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 3, pp. 373-378, March 2003.
18. M. Robeer, G. Lucassen, J. M. E. M. van der Werf, F. Dalpiaz , S. Brinkkemper, "Automated Extraction of Conceptual Models from User Stories via NLP," 2016 IEEE 24th International Requirements Engineering Conference (RE), Beijing, 2016, pp. 196-205.
19. W. Anwar, X. Wang, Lu Li and X. Wang, "A Statistical Based Part of Speech Tagger for Urdu Language," 2007 International Conference on Machine Learning and Cybernetics, Hong Kong, 2007, pp. 3418-3424.
20. https://scholar.google.com/citations?view_op=search_authors
21. https://www.geeksforgeeks.org/supervised-unsupervised-learning/

AUTHOR PROFILE

Kalpana Uppada is currently pursuing her Bachelors at Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. Her areas of research are Semantic web and Machine Learning. She is a member of Computer Society of India (CSI).

Dr. Mandava Kranthi Kiran is working as assistant professor in GITAM Institute of Technology, GITAM Deemed to be University. He has 9 years of teaching experience. He has done his Bachelor's degree from GITAM College of engineering, Masters from BTH, Sweden and attained his PH.D. from GITAM Deemed to be University. Until now he has published upto 12 articles in both international journals and international conferences. He has an experience of 10 years of research and his areas of interests are semantic web, software engineering, artificial intelligence. He is a member of ACM, IEEE and a life member of Computer Society of India.

Sridhar Seepana is currently pursuing his Bachelors Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. His areas of research are Semantic web and Machine Learning

Jahnavi Setti is currently pursuing her Bachelors at Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. Her areas of research are Semantic web and Machine Learning.

Gipson Nikil is currently pursuing his Bachelors at Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. His areas of research are Semantic web and Machine Learning