A SURVEY ON MACHINE LEARNING TECHNIQUES FOR TITLES EXTRACTION FROM REFERENCES

Mr. P Krishnanjaneyulu, U. KALPANA, S. SRIDHAR, S. JAHNAVI, K. GIPSON NIKIL

Assistant Professor, pkrishna.cse@anits.edu.in, Department of CSE.

Anil Neerukonda Institute of Technology & Sciences (Autonomous), Visakhapatnam, AP.

ABSTRACT

Researchers basically collect many research papers inorder to research and study. Collecting numerous data in pdf format might end in redundancy and is additionally space consuming. So, to avoid this problem we have proposed an unsupervised technique to extract the components of references and then identify the title and redirect the title to google search if the author wishes to look that specific title. Till date there are many supervised and regular expression techniques to retrieve the info supported the user query. Though regular expression techniques are good at extracting the titles, if there's any change within the format of the reference, the machine might not identify title correctly because it is trained with only particular formats and there no unsupervised techniques used. So we proposed an unsupervised technique to enhance the performance. We have deeply researched about supervised, unsupervised techniques and various methods implemented to extract titles with the help of unsupervised technique which we want to discuss in this paper.

KEYWORDS: *Machine learning, supervised machine learning, unsupervised machine learning, refernces, regular expressions.*

1. INTRODUCTION

1.1 MACHINE LEARNING

As we all know today, machine learning is transforming the entire world by enabling machines the power to undertake and perform all types of 'intelligent' tasks like understanding images, human speech, predicting preferences etc. [3]. With great deal of knowledge interconnectedness and large processing power in small devices, machines do things which weren't anticipated until recently. While, on the opposite hand, machines are still unable to undertake and perform some tasks which we do easily like image recognition. the method of learning begins with observations or data, like examples, with experience, or instruction, pattern identification in data and make better conclusions of output with the examples that we offer . the most goal of machine learning is to permit the computers learn by self without human involvement or assistance and predict accurate outputs accordingly.

Machine learning algorithms are classified as Supervised and unsupervised. Supervised algorithms require an data scientist or data analyst with machine learning skills to supply both input and desired output supported the training. Data scientists determine which variables, or features, the model should analyze and use for prediction of outcomes. Once training for the machine is completed with the assistance of coaching data, the corresponding algorithm are going to be applied.

Unsupervised algorithms don't undergo any training with data i.e. there'll not be any training data. Instead, they use an iterative approach called deep learning for the analysis of knowledge and to draw conclusions. Unsupervised learning algorithms -- also called neural networks -- are used for more complex processing tasks than supervised algorithms, including image recognition, speech-to-text and tongue generation.

These neural networks work through many samples of coaching data and automatically identifying often subtle correlations between many variables. These algorithms have only become feasible within the age of massive data, as they require massive great deal of coaching data. even as there are nearly limitless uses of machine learning, there's not any shortage of machine learning algorithms. they vary from the fairly simple to the highly complex.

1.2 POPULAR MACHINE LEARNING TECHNIQUES

Two of the foremost widely adopted machine learning methods are supervised learning and unsupervised learning [10], but there are other methods of machine learning Supervised learning algorithms are trained using labelled examples, a touch of kit could have data points labeled either “F” (failed) or “R” (runs). As the the machine is already trained with training data and is now tested with testing data to view and analyze the extracted outputs with the actual outputs to predict out the errors. By this way the machine gains experience to rectify its errors accordingly. Even though there are methods like regression , classification, association, supervised learning uses patterns to assume the values of the label on additional unlabeled data, Supervised learning is mostly observed in applications where historical data predicts likely future events. As an example, it can anticipate when mastercard transactions are likely to be fraudulent or which insurance customer is probably visiting file a claim. Supervised learning models have some advantages over the unsupervised approach, but they even have limitations. The systems are more likely to make judgments that humans can relate to, as an example, because humans have provided the premise for decisions. However, within the case of a retrieval-based method, supervised learning systems have trouble managing new information. If a system with categories for cars and trucks is presented with a bicycle, as an example, it might must be incorrectly lumped in one category or the opposite. If the AI system was generative, however, it's visiting not know what the bicycle is but would be ready to recognize it as belonging to a separate category. Training data for supervised learning includes a gaggle of examples with paired input subjects and desired output (which is additionally said because the supervisory signal). In supervised learning for image processing, as an example, an AI system could be given labelled pictures of vehicles in categories like cars and trucks. After a sufficient amount of observation, the system should be ready to distinguish between and categorize unlabeled images, at which era training are often said to be complete. Unsupervised learning is employed against data that has no historical labels. The system isn't told the "right answer." The algorithm must determine what's being shown. The goal is to explore the knowledge and find some structure within. Unsupervised learning works well on transactional data. as an example, it segments of consumers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find the foremost attributes that separate customer segments from one another. Mostly used unsupervised machine learning techniques involves self-organizing maps, k-nearest neighborhood, k-means clustering and singular value decomposition. These algorithms are wont to segment text topics, recommend items and identify data outliers. Even though there are no categories provided, an AI system can categorize unsorted data according to their similarities and dissimilarities. AI systems capable of unsupervised learning are often associated with generative learning models, although they may also use a retrieval-based approach (which is closely related to supervised machine learning technique). While compared to supervised machine learning techniques, unsupervised machine learning techniques can perform more complex tasks. But, unsupervised learning is more unpredictable as there is no training data. While an unsupervised learning AI system might, for example, identify by itself how to differentiate images such as lions from tigers, it might also add unforeseen and undesired categories to deal with unusual breeds, instead they create a cluster.

2. LITERATURE SURVEY

As we are first to research about unsupervised method of extracting references, we have analyzed different methods of supervised techniques inorder to extract references. Some of them are:

2.1 SEMANTIC TEXT ANALYSIS USING MACHINE LEARNING

User specific suggestions are now becoming a go-to scenario for all recommendation systems. This project proposes such methodology for refining the suggestions provided to a user based on his search history in the software and by extracting the domain from his query[1]. This system is based on semantic text analysis using natural language processing in machine learning and provides a user-specific, personalised suggestions on the user’s queries.

2.2 SODHANA

Sodhana which is our motivation is a semantic file sharing application where the standard client server architecture of the prevailing systems is replaced by the peer-to-peer model of networking[2]. Distributed servers are wont to maintain a non-centralised server and user searches are refined and given optimal suggestions supported the user's query and former searches using machine learning. the need for the usage of the semantic technology has also been stressed by the authors of for the representation of bibliography.

2.3 CERMINE:

Cermine mainly performs two tasks

1. Metadata extraction
2. References extraction

When a research paper in pdf format is passed as an input to cermine, it outputs abstract, keywords, author of the research paper, its title and bibliographic info such as publication year. Cermine uses both supervised and unsupervised techniques. It used support vector machine (SVM) for metadata classification and extraction and K-means clustering algorithm for references extraction. Conditional random fields (CRF) for extracting metadata information from reference strings [9]. The average F-score of cermine is 77.5%.

2.4 GROBID:

This tool is one among the standards adopt by digital library community for parsing bibliography references with respectable performances and high accuracy [4]. These entities identify form scientific article .The automatic extraction of bibliographical data may be a difficulty task. Both scientific article and bibliographies are considered as a structured text, they present a high variability of the formats. Where as this tool comes with more functionalities that perform document's metadata extraction from organising parts of scientific documents text like affiliation extraction, header extraction and bibliography references extraction. Crossref web service for every extracted citation so as to correct the entities extracted from the reference.

2.5 CROSSREF:

Crossref API works differently from cremine and grobib [4]. This tool was employed by Grobid for consolidating its output on the entities extracted from the bibliography. membership association for publishers established in 2000 is not-for-profit. It provides reference linking services for over 62 million scholarly content items like journal articles. by connecting a singular Digital Object Identifier (DOI) to the metadata about the thing, like URL, indicating where the thing are often found. This tool allows querying the Crossref database by giving it in input pattren that contain bibliography references.

3. FEATURES

There are two types of features: format features and linguisticfeatures. We mainly use the former. The features are used for both the title-begin and the title-end classifiers.

3.3.1 Format Features

Font Size: There are four binary features that represent the normalized font size of the unit (recall that a unit has only one type of font).If the font size of the unit is the largest in the document, then the first feature will be 1, otherwise 0. If the font size is the smallest in the document, then the fourth feature will be 1, otherwise 0. If the font size is above the average font size and not the largest in the document, then the second feature will be 1, otherwise 0. If the font size is below the average font size and not the smallest, the third feature will be 1, otherwise 0. It is necessary to conduct normalization on font sizes. For example, in one document the largest font size might be '12pt', while in another the smallest one might be '18pt'.

Boldface: This binary feature represents whether or not the current unit is in boldface.

Alignment: There are four binary features that respectively represent the location of the current unit: 'left', 'center', 'right', and 'unknown alignment'. The following format features with respect to 'context' play an important role in title extraction.

Empty Neighboring Unit: There are two binary features that represent, respectively, whether or not the previous unit and the current unit are blank lines.

Font Size Change: There are two binary features that represent, respectively, whether or not the font size of the previous unit and the font size of the next unit differ from that of the current unit.
Alignment Change: There are two binary features that represent, respectively, whether or not the alignment of the previous unit and the alignment of the next unit differ from that of the current one.
Same Paragraph: There are two binary features that represent, respectively, whether or not the previous unit and the next unit are in the same paragraph as the current unit.
Word Count: A title should not be too long. We heuristically create four intervals: [1, 2], [3, 6], [7, 9] and [9, ∞) and define one feature for each interval. If the number of words in a title falls into an interval, then the corresponding feature will be 1; otherwise 0.
Ending Character: This feature represents whether the unit ends with ':', '-', or other special characters. A title usually does not end with such a character.

4. WORKING MODEL

An research paper in pdf format is given as input. Whole text from pdf is extracted with the help of slate in to a text file. Slate tool helps us in extracting text from pdf format. Only the reference part from the text file is identified and is copied into another text file. As we know, references contain mainly 3 parts. The first one will be author names followed by title of the paper and then publication details like conference or the journal.

4.1 ELEMINATING AUTHOR NAMES AND PROCEEDINGS

Author names will be named nouns. So we have used POS (parts of speech) tagging technique which identifies NNP's (proper nouns) as all named nouns will be proper nouns. So author names will be eliminated. Further keywords such as conference, proceedings, international are identified and the whole text from that keyword is eliminated until a number is encountered. we do this inorder to separate one reference from another as each reference will be starting with a number and also will be ending with publication year or page number which is again a number. So we considered number as a separation between references.

4.2 TITLE EXTRACTION

Title extraction is mainly done with the help of NLP (natural language processing). Using the above process by eleminating author names and proceedings in the references, we will be left over with only titles and numbers (like publication year and page number). Numbers can be eliminated with by using POS tagging again as all numbers will be identified as CD (Coordinating Conjunction). So, all the numbers will also be removed. Finally we will be having only titles in our reference part. Now the author can directly search only desired title with the help of google.

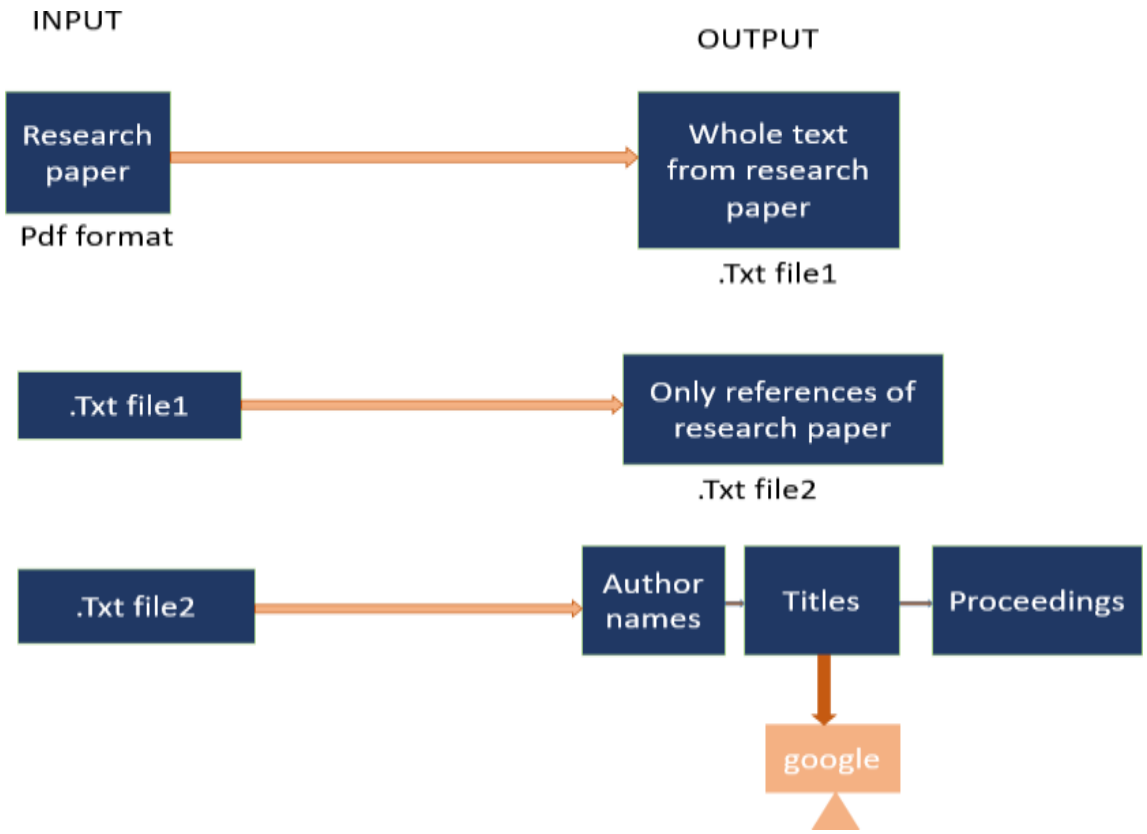

Fig 1: Working model for extraction of titles

5. EXPERIMENTAL RESULTS

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

We have taken different formats of paper to analyze the correct prediction of titles.

Average Accuracy =

$$\frac{\text{Number of exact titles extracted}}{\text{Total number of titles}}$$

Various formats such as IEEE, ACM, SPRINGER etc. resulted us in various accuracies. Taking the average of all the formats Using unsupervised technique resulted us in 72.9% accuracy.

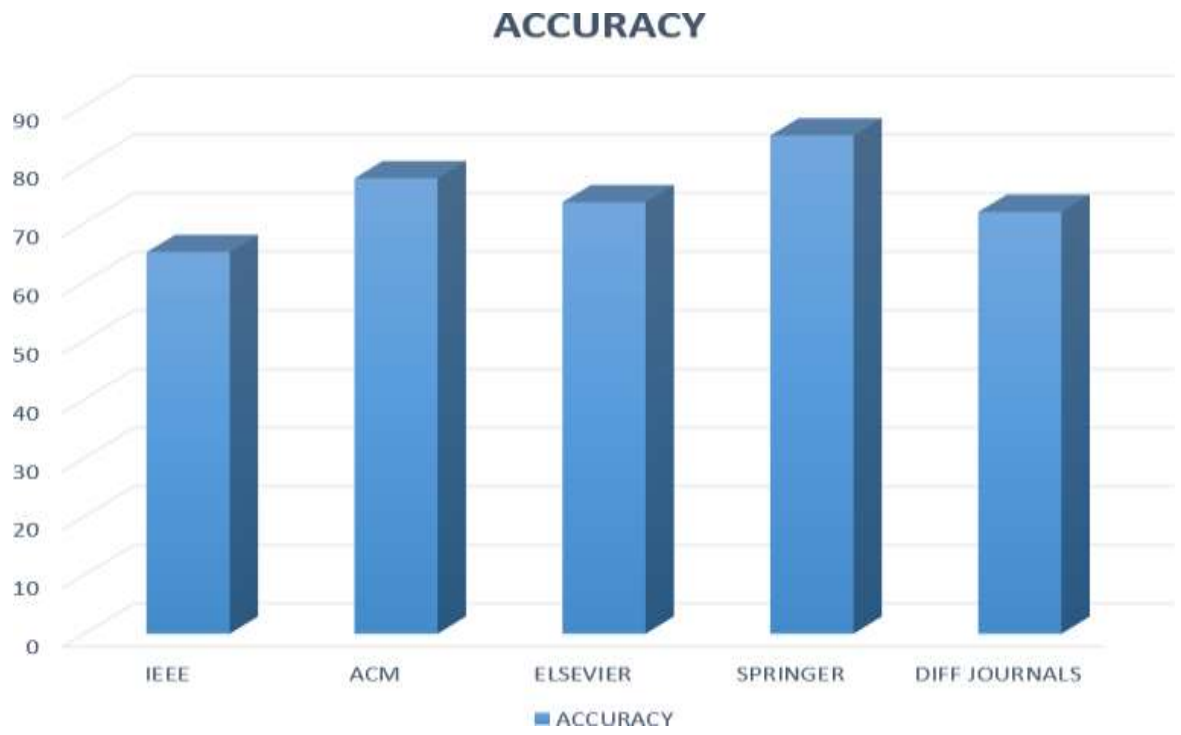

Fig 2: Accuracy analysis for different formats

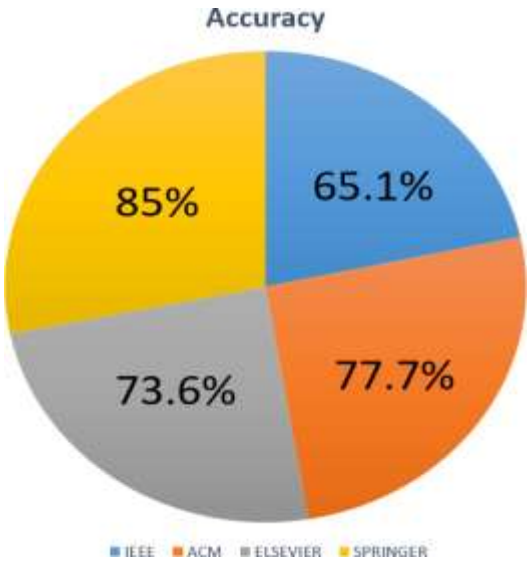

Fig 3: Accuracy analysis of major formats

6. COMPARISION WITH OTHER TECHNIQUES

S.NO	RESEARCH PAPER TITLE	ACCURACY
1	An Approach Towards Establishing Reference Linking in Desktop Reference Manager	78.44%
2	A Strategy for Automatically Extracting References from PDF Documents	74%
3	Unsupervised Technique for Automatically Extracting the Components of References	72.9%
4	Cermine--Automatic Extraction of Metadata and References from Scientific Literature	77.5%

Fig 3: Comparison of accuracies of different techniques for extraction of titles

We have analyzed different techniques for extraction of titles from references. Our main motivation "An Approach Towards Establishing Reference Linking in Desktop Reference Manager" [1] which used java, Xml and RDF triples for extraction of titles resulted in accuracy of 78.44%. "A Strategy for Automatically Extracting References from PDF Documents" which used supervised machine learning technique and regular expressions gave accuracy of 74%. "Cermine--automatic extraction of metadata and references from scientific literature" which again used supervised machine learning technique resulted in accuracy of 77.55% and finally our unsupervised technique for automatically extracting the components of references accuracy is 72.9%.

7. CONCLUSION

Even though some formats resulted in high percentage like springer, as we have calculated average, accuracy is varied. Also we have purely used unsupervised machine learning technique where the machine is not previously trained like in supervised technique. Inspite of using unsupervised machine learning technique, our results are similar to that of supervised machine learning technique as shown in Fig 3. Hence, we strongly assume that using semi supervised technique might increase the performance as it is the combo of both supervised and unsupervised techniques.

REFERENCES

1. Kiran, M. K., & Reddy, K. T. (2018). An Approach Towards Establishing Reference Linking in Desktop Reference Manager. Journal of Information & Knowledge Management, 17(03), 1850034.
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. Joga, B., Sattiraju, S., Kandula, V., Kallempudi, N. M., & Kiran, M. K. (2019). Semantic text analysis using machine learning.
4. Chenet, M. (2017). Identify and extract entities from bibliography references in a free text (Master's thesis, University of Twente).
5. Alves, N. F., Lins, R. D., & Lencastre, M. (2012, March). A strategy for automatically extracting references from PDF documents. In *2012 10th IAPR International Workshop on Document Analysis Systems* (pp. 435-439). IEEE.
6. Turney, P. D., & Littman, M. L. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. arXiv preprint cs/0212012.
7. Figueiredo, M. A. T., & Jain, A. K. (2002). Unsupervised learning of finite mixture models. IEEE Transactions on pattern analysis and machine intelligence, 24(3), 381-396.
8. Arora, C., Sabetzadeh, M., Goknil, A., Briand, L. C., & Zimmer, F. (2015, August). Change impact analysis for natural language requirements: An NLP approach. In 2015 IEEE 23rd International Requirements Engineering Conference (RE) (pp. 6-15). IEEE.
9. Tkaczyk, D., Szostek, P., Dendek, P. J., Fedoryszak, M., & Bolikowski, L. (2014, April). Cermine--automatic extraction of metadata and references from scientific literature. In 2014 11th IAPR International

Workshop on Document Analysis Systems (pp. 217-221). IEEE.
10. 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.
11. https://scholar.google.com/citations?view_op=search_authors
12. https://www.geeksforgeeks.org/supervised-unsupervised-learning/