Knowledge Extraction from Bangla Documents using NLP: A Case Study

Monika Gope and M.M.A. Hashem
Department of Computer Science and Engineering
Khulna University of Engineering & Technology (KUET)
Khulna 9203, Bangladesh
monikagope@iict.kuet.ac.bd, hashem@cse.kuet.ac.bd

***Abstract*—In this paper, we have proposed a system that determines and extracts the user query from the vast store of official Bangla digital documents, and performs detection, and analysis of the documents. A set of different rules and knowledge based methods is used to extract the required decisions from the resolutions of Bangla documents of a specific domain. Then, with diverse parameters, the effects of the process are discoursed and classified and compared with exact match and sematics features of the words with relation with other reative sentences where it achieved a higher performance for a sample dataset with a knowledge base of the documents.**

Keywords—Knowledge Extraction, Natural Language Processing, Bangla Documents, Query, Keywords

I. INTRODUCTION

The information extraction in the digital Bangla documents is a cumbersome process as this usually ends up with incorrect and a very few useful information. The process is difficult because wide computational resources for Bangla language are quite limited till date. The limited computational resources could be attributed to the difficulty in analysis and knowledge extraction of Bangla documents. However, given the widespread use of Bangla language in different important government documents and in digital platforms, it is high time we developed the necessary computational resources.

In our previous work [1] we proposed a technique that discovers and extracts various important decisions from the official Bangla documents. For this purpose, we have used the minutes of meeting of the academic council of Khulna University of Engineering & Technology (KUET) as a domain where data are stored as Bangla PDFs. With a set of regular expressions, the decisions and agendas are recognized inspired by Bhatia, et al. [2], [11] and then the result is shown with the classification model. However, the searching methods used various features and regular expressions and did not use natural language processing or keywords searching policies. As a result, it only fetched the exact data and was not able to find the related knowledge about the user query. In order to fetch the extact and proper information with related data, there should be some classification category of huge data to arrange and analysis.

This paper has the following vital contributions:

1) We have proposed composite approaches to find out the decisions from the decision pool from the meeting resolutions of the Academic Council of KUET based on various Content-based, Semantics-based and Context-based features of the sentences.

2) We have also revealed the techniques to categorize and classify the decisions from the documents using keywords extraction.

The rest of the organization of the paper is as follows: Section II discusses the leading research on information and extraction related to our work. The proposed system is explained in Section III. The experimental outcomes with discussions are clarified in Section IV. Section V concludes the paper.

II. RELATED WORK

There exist several ways for extracting knowledge, such as, figures, tables and useful information from e-documents. The exraction of mathematical expressions [3], [4], tables [5], figures [6], [7], and tables of contents [8] from e-documents have been proposed earlier. To extract text, Kataria, et al. [7], worked with image processing and Bhatia, et al. offered identification of captions and detection of tables, figures, and pseudocodes [9].

However, Information or Keywords related to information search required an authoritative technique, which can enable reading large documents without having any knowledge about the syntax or understanding the semantics of the language [10]. Moreover, a sentence is to be connected to more than one subject expressed within a document or set of documents [12]. One of the pioneer works in this field was presented in [21],[22] where a supervised approach was employed to extract keywords from titles and abstracts of journals and later unsupervised method called TextRank[15] RAKE[21], TAKE[22] are widly used for extraction keywords.

For Bangla language, joint correlation technique is used for Bangla number extraction and recognition from the document image [13] and for knowledge extraction in Bangla documents for emotion or opinion detection from Bangla blogs and news [14] is also explored. But these will not serve our purpose to extraxt knowlwge from the Bangla decision pool of our domain.

III. THE PROPOSED SYSTEM

In this section, the process of keywords extraction and user query extraction from the resolutions of the Bangla document are presented. Figure 1. illustratrates the strategy of the proposed system. First of all, we started to extract the keywords from the processed Bangla documents [1] where

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

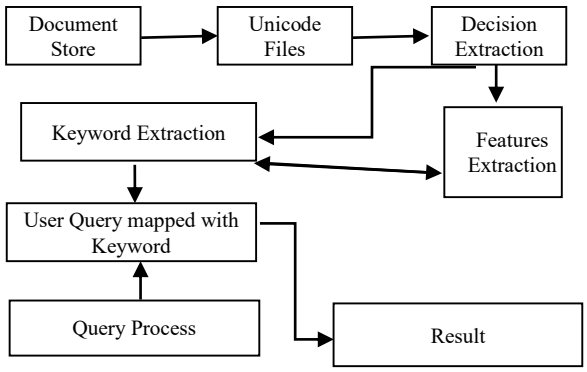

Fig. 1. The proposed system

the decisions are stored. Then, the process of extracting the query knowledge from these decisions is performed. To extract the desired knowledge, we processed the decision store to get the important texts. Then, with the corresponding features, we extracted the knowledge from the text, which is mapped with the user query. Figure 2. presents the proposed algorithm.

The system takes the documents De_i, a set of Bangla decisions and Query Q and returns de_i ,a set of Bangla detected texts. For each query, the system collects the required lines with some semantics features explained in the following sub sections. We divided the process into the major following parts. The proposed model is specified below-

A. Processing of the keywords from the Text

In [1] and in Bhatia, et al., [10], the extracted set of decisions is used. Then, the sentences are cleaned by removing stop words. Here the domains specific stop words are embedded with common stop words. Then, keywords extraction is done with different matured algorithms.

1) There are some common stop words in the text, such as "অথবা", "অনুযায়ী", "অনেক", "অনেকে", "অন্তত", "অন্য" etc. We have used 387 common bangla stopwords and listed 66 domain specific stopwords to the main stopwords list. Some examples are given in the Table I.

2) After the elimination of stop words, from the list of the decision, the term frequency-inverse document frequency, tf–idf [11] or TF-IDF, which reflects important word in a document or collection, is used to extract the keywords. Term frequency for the document is:

$$tf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (1)$$

Where d is the document, t is the term which means the number of times it occurs in d and $\sum_{t' \in d} f_{t',d}$ is the total number of words in d.

The inverse document frequency is a measure of how much information the word provides [11],

$$idf(t, D) = \log \frac{N}{N^t} \quad (2)$$

Where $N = |D|$ is the total number of documents and N^t is the number of documents, where the term t appears.

Algorithm: Decision Extraction with features

Input Decision: $De_i = \{De_1, De_2, D_3, \dots De_n\}$

Query $= Q$ *and Output:* $de_i = \{de_1, de_2, de_3, \dots de_n\}$

1. *For each* $D \in De_i$
   a. $dk \leftarrow Extract\ keyword$
   b. end
2. $dk \leftarrow rank\ keyword$
3. *For each* $S \in De_i$
   a. $Create\ feautures\ set\ with\ keyword$
   b. $S1 \leftarrow Match\ Feauture\ set$
   c. $S2 \leftarrow Match\ Q$
   d. $S \leftarrow Combine\ S1, S2,$ $and\ remove\ duplicates\ if\ any$
   e. end
   c. $S_i = \{S_1, S_2, S_2, \dots S_n\}$
   d. $dei = Si$
   e. end
4. $Order\ (dei\ )$
5. $Return\ dei$

Fig. 2. The Proposed Algorithm

And the *tf–idf* is [11],

$$tf - idf(t, d, D) = tf_{t,d}.idf(t, D) \quad (3)$$

Rapid Automatic Keyword Extraction (RAKE) can spontaneously extract keywords from documents [17]. We have used Multilingual Rapid Automatic Keyword Extraction (MRAKE) [16] for the purpose for detecting the keywords. If the entire documents is used in MRAKE, its stopwords will be more specific as it is language independent, and the stopwords are generated from the text itself. The more texts, the more stopwords.[16].

Textrank described in [15] and [12] creates a graph of the words and relationships between them from a document. Then it finds the most important points of the words based on importance scores calculated from the entire word graph. We have used this algorithm to rank the sentences here.

3) Then, we have selected the most common 20 keywords using the above techniques. The most frequently extracted keywords are listed with the tf-idf and MRAKE score. Two examples are given in Table II.

B. Feature Extraction for Decisions with User Query

In the files, there are many keywords and we have selected the most common and high-frequency keywords from the keyword list as keywords knowledge base. If a user gives a query from these decision files, at first it looks into the keywords knowledge base. The keywords knowledge base

TABLE I. STOPWORDS FOR THE DOMAIN

কার্যক্রম	করেন	শিক্ষার্থী	অন্যান্য	দেওয়া	ডিগ্রী
সকলে	Eng	শিক্ষার্থীর	অনুষ্ঠিত	যাবে	হোক
সকল	ডিগ্রি	Roll	format	পর্যন্তত	ভর্তি
সভায়	মোসা	ধন্যবাদ	১ম	চালিয়ে	তালিকা
সভা	রয়েছে	বিভিন্ন	পর্যন্তত	যাওয়ার	হবে
গৃহীত	গঠন	ভুক্ত	তারিখ	ইহা	পর্যায়ে
উপলক্ষ্যে	সদস্যবৃন্দ	অনুষ্ঠিত	ছাত্রী	ছাত্র	পর্যায়ের
উপলক্ষ্য	শিক্ষার্থীকে	No.	জ্ঞাপন	হলো	in
কোর্স	কোর্সসমূহ	অনুমোদন	গ্রহণের	সুপারিশ	বিভাগ

TABLE II. EXAMPLE OF THE FRQUENCY OF TWO WORDS :POST FACTO AND MECHANICAL

Word	Score IfIDF	Frequency IfIDF	Score MRake	Frequency MRake
post-facto	0.00462, 0.00463	2	1, 4, 4, 4, 4	5
Facto	0.036	1	-	-
ঘটনাত্তর	0.0204, 0.00327	2	-	-
যন্ত্রকৌশল	-	-	1, 1, 1, 1,1	5
মেকানিক্যাল	0.00711	1	-	-
ME	0.01452, 0.00732	2	-	-

has three features specifications on it. If the query is not there in the knowledge base, it will directly search for the matching words. The three features where a phrase or keyword is matched with the query word in a sentences is described below. If we are not able to find the exact information we are looking for [9], [10] in the direct search, we use the other features. The features are discussed below:

1) Content-Based Features: The phrase or user query is directly searched with a set of a regular expression described in [1]. All the lines which contain the phrase are fetched by the regular expression (regex).

However, through this means, we only found out the words that exactly matched the phrase and we did not get relevant information regarding our search [1]. In this case, we have used a set of words called "pratyay" with the query word from the keywords knowledge base. If the query word is "কমিটি" and the word is from the knowledge base there is a set of words with the keyword, such as "কমিটিসমূহ" , "কমিটি-এর " , "কমিটিতে " etc. This is the case for knowlege base word.

2) Semantics Features: In [23] combination of lexical taxonomy with corpus statistical information is used to learn the semantic distance between nodes. Inspired by these [23], we found there are some certain words with various kind of synonyms or semantics similarites which is used frequently to express the same meaning. We find five kinds of synonyms as-

a) In the documents, we found that many words in english are written in bangla font, such as "Sessional" as "*সেশনাল*". We have tried to make a knowledge base with english word from bangla font words. For example, if the user query is "*সেশনাল*", then it can also find word "Sessional". In the 20 keyword knowledge base, we have listed the commonly used English words writted in Bangla font in this domian, such as, "*অর্ডিন্যান্স*" , "*গ্রাজুয়েট*" , "*মেকানিক্যাল*" , "*রেজিস্ট্রেশন*" , "*কমিটি*", "*থিওরী*" etc.

b) Then, there are some English to Bangla words that are used simentaiously, such as "তত্ত্বীয়" with theory, "কোর্স প্রত্যাহার" with course withdrawal, committee with "কমিটি", "ঘটনাত্তর" with Post Facto, "রেজিস্ট্রেশন" with registration, "মেকানিক্যাল" with Mechanical. They are used vise versa in all the documents. Therfore, we also consider this.

c) There are some words with short or elaborated forms, such as, "মেকানিক্যাল" with ME or "এমই", CASR with "উচ্চ শিক্ষা ও গবেষণা কমিটি" etc. We have listed these types of most common keywords for searching.

d) A lot of words are there with Bangla synonyms, such as "নিম্নলিখিত", "নিম্নে", "নিম্নোক্ত" , "নিম্নবর্ণিত", "নিম্নরূপ". All have the same meaning but used randomly throughout the douments. Morover, Bangla words sometimes have different spellings for the same word such as "ঘটনাত্তর" and "ঘটনাত্তোর" etc. For the 20 keywords, we have made these knowledge base.

e) We have also classified some words which shows the decision is very significant and urgent. Some of such words are "জরুরী", "অধিকতর", "দ্রুত", "সতর্ক", "জরিমানা", "Compulsory", "স্থগিত", "বহুল", "অঙ্গীকারনামা" , "যতশীঘ্র" , "কঠোরভাবে", "শীঘ্র" ,"গুরুত্বপূর্ণ" and therefore are added in the knowledge base.

The Bangla wordnet [18] can be used for Bangla similar words but we can see from the above discussuion that there are many English word wrritten in Bangla font.Therefore, only Bangla to Bangla meaning and English to English meaning extraction will not give an accurate result. So far, we have applied many semantic conditions and did not consider the surroundings clues of the sentence. Hence, we need Context-based features to find out the exact context of the knowledge.

3) Context-Based Features: In our previous work [1], we did not include natural language processing and, we have extracted the knowledge only by the location of the sentence. Here we have considered the context and its connection to the required sentences. Some sentences refer to another sentence as a reference or the sentences are connected to a definite topic. If we don't find the connected lines with the query, then the we will miss the important information from the result and it will not be the exact knowledge. From all the decision files in our working domain from the data set, we found five types of connecting words which are illustrated in Table III and discussed below:

a) A connection word list that indicates the specific topics previously described.

b) A connection word list that indicates person/s after a sentence.

c) A word list of conditionally connecting words immediately after the sentence used in the dataset.

d) A list of words that immediately talks about near future or past with the sentence.

e) Explanation words that expains the topics about the previous line.

TABLE III. CONNECTION WORD LIST

Immediate Definite Topics	Indicating person	Adding extra Condition	Steps in future or past	Explain Immediate line
উল্লিখিত	তাকে	তবে	ভবিষ্যতে	অর্থাৎ
উক্ত	তাঁকে	এছাড়া	পরবর্তী	
এ ব্যাপারে	তাদের	উভয়ক্ষেত্রে	পরবর্তীতে	
অত্র	তাঁদের	অন্যথায়	অতঃপর	
ইহা	তাঁদেরকে	এছাড়াও	ইতঃপূর্বে	
উপরোক্ত	তাঁর	এতদসাথে		
এই		অপর		
এটাকে				

If the query word is found in the sentence, these words are immediately searched in the consecutive next two sentences. However, if the query words are found in these connection sentences, then the previous sentence is extracted with the current sentence. These connection words are mostly the first word of the sentence. However, they can be anywhere in the sentence according to the context. So we have considered these words as location independent in the sentence.

C. Categorization of the Documents with Keywords:

All the files with the decision are categorized with the 20 keywords and stored in the files. If the user query is related to these keywords then the information stored in the files is used to extract the knowledge. From the set of total 29 documents, we have extracted 698 decision making lines. From these 698 lines, we have extracted most frequent 20 keywords with tf-idf and MRAKE. Then with these 20 keywords all the feature extraction is done as described above. Then we have categorized 20 different topics with 20 keywords.

IV. Experiments and Result Analysis

For result analysis, we have experimented with 29 meeting resolutions of Academic Council of KUET to extract the data and then using the method [1], collected the decision list. We assumed all the words and lines are correctly extracted, otherwise keywords and decisions may not correctly found. From these, we have applied the described methodology. It is divided into three parts: *"Keywords* detection, *"Query"* detection, and categorization.

All experiments in done on a Windows machine and we used Python as ~~a~~ the programing language to implement our algorithm and Weka [19] as the implementation tool for classification. We have used nltk and other python packages. Naïve-Bayes Classifier model and Gaussian techniques are used to classify the result.

We have also measured the cosine similarity of the sentences and rank them with text rank algorithm [15], [20] to get the most informative sentences. For 698 decision sentences, we have made a similarity matrix to know the top relevant sentences. In Figure 3. it showed the top textrank sentences using cosine similarity where the red words are the keyword extracted by tf-idf and MRAKE algorithm. Most of these words are from our 20 keyword knowledge base. Textrank also verifies the common frequency words from both tf-idf and MRAKE. From the textrank, we ranked all the 698 lines and the values are then sent to the Gaussian model to make a classification. We

1. সিদ্ধান্তঃ .....বিশ্ববিদ্যালয়ের পুরকৌশল, তওইকৌশল, যন্ত্রকৌশল, সিএসই ও ইসিই বিভাগের বিভাগীয় প্রধানগণ কর্তৃক প্রস্তাবিত এবং Co-ordination কমিটি কর্তৃক সুপারিশকৃত একাডেমিক ক্যালেন্ডারসমূহে নির্ধারিত ছুটিগুলো ঠিক রেখে তৈরী করার জন্য পরামর্শ দেয়া হয়।

2. একাডেমিক ক্যালেন্ডারসমূহ সংশোধিত আকারে আবারো অনুমোদনের জন্য আনা হোক।

3. সিদ্ধান্তঃ ..... বিশ্ববিদ্যালয়ের পুরকৌশল, তওইকৌশল, সিএসই ও ইসিই বিভাগের প্রস্তাবিত বিভিন্ন বর্ষের নিয়মিত পরীক্ষা কমিটি (পরিশিষ্ট-জ্ঞ০৭/০৪ঞ দ্রষ্টব্য) অনুমোদন করা হলো।

4. সিদ্ধান্তঃ ..... বিশ্ববিদ্যালয়ের তওইকৌশল ও সিএসই বিভাগের স্পেশাল ব্যাকলগ পরীক্ষার জন্য গঠিত পরীক্ষা কমিটি অনুমোদন করা হলো (পরিশিষ্ট-জ্ঞ০৭/০৫ঞ দ্রষ্টব্য) ।

5. সিদ্ধান্তঃ ..... বিশ্ববিদ্যালয়ের শিক্ষাবর্ষের পুরকৌশল, তওইকৌশল, যন্ত্রকৌশল ও সিএসই বিভাগের বিভিন্ন বর্ষের বি. এস-সি. ইঞ্জিনিয়ারিং পরীক্ষার ফলাফল (পরিশিষ্ট-জ্ঞ০৭/০৬ঞ) ও ডিগ্রী অনুমোদন করা হলো।

Fig. 3. The Top-ranked sentences by Textrank

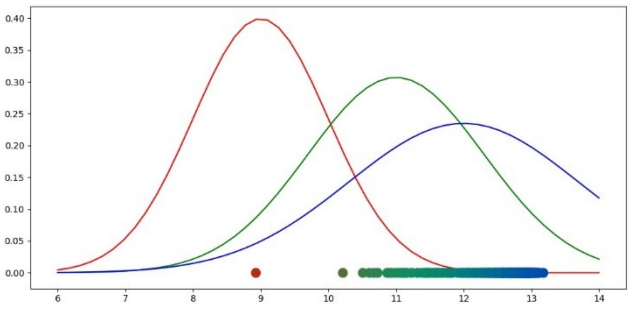

Fig. 4. Gaussian curve for three cluster

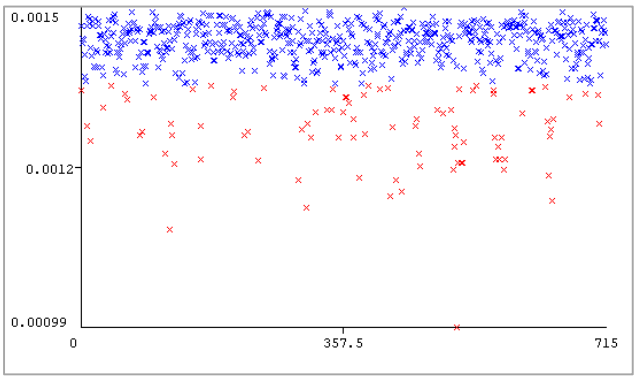

Fig. 5. k-means curve of the three cluster

have made three clusters to represent the graph which is shown below in Figure 4. Here we can understand the blue and green groups are overlapping with each other. Therefore, there are many lines which holds the same words as keywords hence share same cluster. And the cosine similarities of the words are also very close to each other as it is a specific domain and most of the topics are expressed with similar words. However, for K-means clustering curve for 698 lines, it made two-clusters in an unsupervised learning that is showed in Figure 5. It also concludes that these sentences are very similar as shown in Figure 3.

Then by the model, we have extracted the desired information for different keywords with three features. For these three features, 10-fold cross-validation is used for precision, recall, and F1 measures to evaluate the result. They are defined as [2]:

$$Precision: = \frac{\text{set of data to be detected} \cap \text{set of detected data}}{\text{set of detected data}}$$

$$Recall: = \frac{\text{set of data to be detected} \cap \text{set of detected data}}{\text{set of detected detected}}$$

$$F1 = \frac{2.Precision \times Recall}{Recall + Precision} \quad (5)$$

In Table IV, we have shown the precision, recall, and F1 for 698 decision from the set of 29 documents for a single query "মেকানিক্যাল" where high precision and high recall are found which means the result is fetched appropriately. Here, we find the recall are 1. And the precision is 92% with 96% F1 measure. And the accuracy [10] is:

$$Accuracy = Detected\ sentences / Total\ sentences \quad (6)$$

TABLE IV. TOTAL SET OF DATA FOR ONE KEYWORD – "মেকানিক্যাল" WITH PRECISION AND RECALL AND F1

Total number of decision lines	698	Pr	Rc	F1
Total number of 'মেকানিক্যাল' related lines from decisions	93			
Total number of detected only 'মেকানিক্যাল' lines from decisions [1]	02			
Total number of detected 'মেকানিক্যাল' related lines from decisions	87	.925	1	.961
Total number of documents	29			

In [1], from 29 documents the query of "*Mechanical*" as "*মেকানিক্যাল*" is explored and only in a single document the query is found in 16 lines. In these 16 lines there are only 2 lines in decision sentences in a document [1]. But using the described three features, here we found 87 corresponding lines with the related words with "*মেকানিক্যাল*" from all the decision pool. Here, the knowledge base set was "যন্ত্রকৌশল", "যন্ত্রকৌশল অনুষদ", "যন্ত্রকৌশল বিভাগ", "ME", "মেকানিক্যাল", "Mechanical", "mechanical", "এমই". For the seven query words the number of lines extracted by the proposed methods are shown in Table V with exact match. Here we found that the proposed system showed higher performance than a exact search [1] for a sample data set with a set of keywords knowledge base.

TABLE V. NUMBERS OF LINES FOUND WITH FEATURES AND EXACT MATCHING

Query words	Number of lines found for Keywords Set with Connecting Word list	Number of lines found for exact matching
Post-facto	23	7
গ্রাজুয়েট	25	8
পরীক্ষক	140	18
ধারা	38	27
সংশোধিত	82	8
Ordinance	22	15
তত্ত্বীয়	13	2

V. CONCLUSION

Knowledge Extraction from a set of Bangla files is very significant for decision making. In [1], we presented a model, extraction system for Bangla official documents. However, for the query searching, direct word matching is used which could not extract the important information related with the query. In this work, we have presented the features with natural languages processing where semamtics and context of the sentences are used. Here results are presented with a precision and recall and exhibited that the proposed algorithm attained a higher performance for a set of keywords knowledge base. The accuracy for these particular knowledge base on a sample dataset is more than 90%. However, the major disadvantage of the process is that the keywords knowledge base is small and do not extract information using vocabulary and morphological investigation of words.

REFERENCES

[1] M.Gope, M.M.A. Hashem, Knowledge Extraction from Bangla Documents: A Case Study, in Conf. on Bangla Speech and Language Processing (ICBSLP), 2018.

[2] S. Bhatia and P. Mitra, Summarizing figures, tables, and algorithms in scientific publications to augment search results in Journal on Information Systems, TOIS 2012, 30(1), 3:1–3:24.

[3] J. B. Baker, A. P. Sexton, V. Sorge, and M. Suzuki, Comparing approaches to mathematical document analysis from pdf in Conf. on Document Analysis and Recognition, ICDAR 2011, pp. 463-467.

[4] R. Zanibbi and D. Blostein, Recognition and retrieval of mathematical expressions in Journal on Document Analysis and Recognition, IJDAR 2012, 15(4), pp. 331-357.

[5] S. Mandal, S. P. Chowdhury, A. K. Das, and B. Chanda, Automated detection and segmentation of table of contents page from document images in Conf. on Document Analysis and Recognition, ICDAR 2003, pp. 398-402.

[6] P. Chiu, F. Chen, and L. Denoue, Picture detection in document page images, in Conf. on document engineering, DocEng 2010, pp. 211-214.

[7] S. Kataria, W. Browuer, P. Mitra, and C. L. Giles, Automatic extraction of data points and text blocks from 2-dimensional plots in digital documents in Conf. on Artificial intelligence, AAAI 2008, Vol 2 pp. 1169–1174.

[8] Z. Wu, S. Das, Z. Li, P. Mitra, and C. L. Giles, Searching online book documents and analyzing book citations in Conf. Document engineering, DocEng 2013, pp. 81–90.

[9] P.W. Staar, M. Dolfi, C. Auer, and C. Bekas, Corpus Conversion Service: A machine learning platform to ingest documents at scale in Conf. on Knowledge Discovery & Data Mining, KDD 2018, pp. 774-782.

[10] H.M. Lynn, E. Lee, C. Choi and P. Kim, Swiftrank: an unsupervised statistical approach of keyword and salient sentence extraction for individual documents, in Procedia of Computer Science, 2017, 113, pp.472-477.

[11] S. Tuarob, S. Bhatia, P. Mitra and C. L. Giles, AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data, in Journal on Big Data, 2016, 2(1), pp. 3-17.

[12] A. Skabar and K. Abdalgader, Clustering sentence-level text using a novel fuzzy relational clustering algorithm, in Journal of knowledge and data engineering, 2013, 25(1), pp.62-75.

[13] M. K. I.Molla, and K. M.Talukder, Bangla number extraction and recognition from the document image in Conf. on Computer and Information Technology, ICCIT 2007, pp. 512-517.

[14] A. Das and S. Bandyopadhyay, Phrase-level Polarity Identification for Bengali, In Journal of Computational Linguistics and Applications, IJCLA 2010, 1(1-2), pp. 169-182.

[15] R. Mihalcea, Graph-based ranking algorithms for sentence extraction, applied to text summarization in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Association for Computational Linguistics, 2004, p. 20.

[16] https://github.com/vgrabovets/multi_rake [Accessed on Nov 10, 2018]

[17] S. Rose, D. Engel, N. Cramer and W. Cowley, Automatic keyword extraction from individual documents, In Journal of Text Mining: Applications and Theory, 2010, pp.1-20.

[18] http://indradhanush.unigoa.ac.in/public/webcontent/webcontent.php?id=37, [Accessed on Nov 10, 2018]

[19] Weka, https://www.cs.waikato.ac.nz/ml/weka/, [Accessed on Nov 10, 2018]

[20] T. Pay, S. Lucci and J.L. Cox, An Ensemble of Automatic Keyphrase Extractors: TextRank, RAKE and TAKE, 2018

[21] A. Hulth, Improved automatic keyword extraction given more linguistic knowledge, In Conf. on Empirical methods in natural language processing Association for Computational Linguistics, 2003, pp. 216-223.

[22] T. Pay, Totally automated keyword extraction, In Conf. on Big Data 2016, pp. 3859-3863

[23] J.J. Jiang and D.W. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, In Conf. of International Conference Research on Computational Linguistics, 1997, arXiv preprint cmp-lg/9709008. 1997