Information Extraction from Research Papers Based on Statistical Methods

Selvani Deepthi Kavila and D. Fathima Rani

Abstract In the research field we require more time to read a single research paper; it also consumes more time to find the algorithms and limitations of the paper. So we require a fast reading system to identify this problem. This paper identifies the algorithms or techniques, or methods, and limitations of a research paper, and also classifies the area of the algorithm. Key phrases are sets of words that elucidate the relationship between context and content in the document. Key phrases are identified from the document and algorithms or techniques, or methods of that paper are extracted. Keywords are a subset of words that contain important information and the area is classified. Cue words are those that contain meaningful information used to identify the limitations of the paper.

Keywords Text mining · Information extraction · Key phrase extraction · Key words · Cue words

1 Introduction

In computer science, text mining has become an important research area. The process of getting extremely useful information from an unstructured text is known as text mining. Text mining is same as data mining, except the tools designed in data mining are used to handle structured data from databases, but text mining can work with semi-structured data, where as HTML files, emails, and full-text

S.D. Kavila · D.F. Rani (✉)
Department of Computer Science and Engineering, Anil Neerukonda Institute
of Technology and Sciences, Sangivalasa, Bheemunipatnam (M), Visakhapatnam, India
e-mail: fathimadirisina@gmail.com

S.D. Kavila
e-mail: selvanideepthi.cse@anits.edu.in

© Springer India 2016

S.C. Satapathy et al. (eds.), *Proceedings of the Second International Conference on Computer and Communication Technologies*, Advances in Intelligent Systems and Computing 381, DOI 10.1007/978-81-322-2526-3_59

documents, etc, and also includes unstructured text. Information extraction is one of the technologies that have been developed and can be used in the text mining process. In text mining, information extraction is an important research area. One of the text mining techniques [1, 2] is information extraction which means extracting of structured information from unstructured documents and semi-structured documents. In many cases, this action concerns the processing of human language text by means of Natural Language Processing (NLP). Text mining has strong capable connections with NLP.NLP has originated from technology that trains computer natural language so that they may examine, understand, and even generate text

Information Extraction is the initial point for computers to examine unstructured text. Information extraction software identifies key phrases and relationships present within the text. It happens by pattern matching, which is a process of exploring for predefined sequences in text. The software interprets the relationship among the identified people, time, and places, where it provides significant information to the user. For dealing with large or huge volumes of text this technology can be very useful.

In this paper, we describe a framework that can extract algorithms or methods or techniques from research papers, and classify text of research documents arrived to the repository and extract demerits of the research paper. The proposed framework consists of two phases; for module 1 first extraction of key phrases [3, 4] and for module 2 it extracts keywords for classification of area [5] and categorization of algorithms, and for module 3 it extracts cue words from the documents, where first it constructs the list of key phrases, keywords, and cue words for the papers and in the second phase it updates key phrases, keywords, and cue words list associated with the new stream of research documents.

The remaining sections of the paper are as follows: Sect. 2 discusses the related work based on text mining, key phrase extraction using information extraction technique, and classification using area tracking, cue words based text extraction, Sect. 3 describes the problem statement related to stream of documents dynamically added to repository. The proposed architecture and framework explained in Sect. 4 whereas experimental results are reported in Sect. 5, and finally we conclude the paper in Sect. 6.

2 Related Work

"*Text mining*" demonstrates the automated discovery of useful or interesting knowledge from unstructured text by application of data mining techniques [1, 2]. There are several techniques proposed for text mining, which include conceptual structure, episode rule mining, decision trees, association rule mining, and rule induction methods. In addition, information retrieval (IR) techniques are widely used as the "bag-of-words" model [2] for tasks such as document matching, ranking, and clustering. Mooney and Nahm [5] proposed a DISCOTEX (Discovery from Text Extraction) approach. For extracting a structured database from a text corpus, this approach uses an automatic learned information extraction system. Then it mines this database with existing KDD tools. The technique was applied to a corpus of computer job announcement postings from an Internet newsgroup. IE has shown to be useful in various other applications such as news articles on corporate acquisitions, seminar announcements, university web pages, apartment rental ads, and restaurant guides [5].

Zhang et al. [6] proposed an automatic keyword extraction from documents using conditional random fields . Conditional random fields (CRF) model is a cutting edge grouping naming technique, which can utilize the peculiarities of records more sufficiently and adequately. In the meantime, decisive words extraction can be considered as the string naming. In this paper, we propose and execute the CRF-based watchword extraction approach. Trial results demonstrate that the CRF model beats the other machine learning routines, for example, bolster vector machine, various direct relapse model, and so forth in the undertaking of watchwords extraction from scholastic reports.

Menaka and Radhika [7] proposed a “Text Classification using Keyword Extraction Technique” content grouping as one of the real utilizations of machine learning. They proposed a technique to use content mining calculations to concentrate watchwords from research papers. The extricated pivotal words have the most astounding similitude. At that point, archives are arranged taking into account separated words utilizing the machine learning calculations—k-nearest neighbor, naïve Bayes, and decision tree. The execution of machine learning calculations for concept of characterization demonstrates that the decision tree calculation gives better results in view of forecast exactness when contrasted with other calculations.

Pinaki bhaskar Kishorjit nongmeikapam Sivaji bandyopadhyay [8] proposed “Key phrase Extraction In Scientific Articles: A Supervised Approach.” This paper contains the definite methodology of programmed extraction of key phrases from logical articles (i.e., examination paper) using regulated instrument like conditional random fields (CRF).

Wan and Xiao [4] proposed single document key phrase extraction using neighborhood knowledge. This paper affirms to utilize a little number of closest neighbor archives to give more learning to enhance a single report key phrase extraction

Learning to extract key phrases [9] from Text P. Turney In this paper, they approach the issue of consequently extricating key phrases from the content as a managed learning undertaking.

Hanyurwimfura et al. [10] proposes a method for cue word-based text extraction.

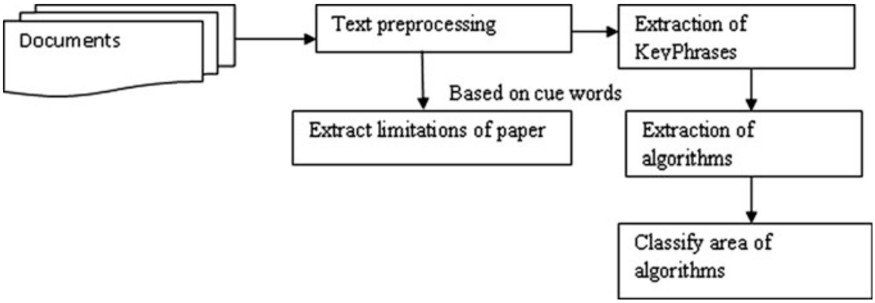
3 Proposed Work

A text document is an ordered sequence of words that arrives in timely order. Document streams are different from documents stored in static repository. They are uninterrupted, limitless, and come at high speed. Let R be the repository of d data mining and m be the maximum number of key phrases. Let St *n* be stream of *n* documents arriving at time stamp t, and Fnxm be feature word matrix associated with recently added documents. Let Lnxm be feature word list for d areas and each area has maximum of m feature words.

Step 1: Read the text from document
Step 2: Preprocess the text.
Step 3: Remove the stop words from document.
Step 4: Construct the list of key phrases, keywords, and cue words and their counts.
Step 5: Extract keywords, key phrases from the document and match with the feature list

4 Proposed Architecture

The above figure represents the architecture in which the stream documents arrive, then each document applied to sequence of activities preprocessing, key phrase, keyword extraction, and cue words are identified. The documents are moved to repository one by one and storage management is based on processing models like Land mark, sliding window model, and Damped model.



4.1 *Stages of Proposed Work*

4.1.1 Information Extraction from Document

The first step for the system using information extraction is to classify the unstructured text. It is a software process of identifying key sentences, key phrases, and relationships within the document. It uses predefined sequences in document called pattern matching.

4.1.2 Key Phrase Extraction and Keyword Extraction

Key phrase extraction method contains three steps: document preprocessing, noun phrase identification, and key phrase extraction.

(**a**). The preprocessing task includes formatting of each document. The pdf document is converted into text format.

Documents preprocessing is to be done

Stop word elimination: Stop words are a part of natural language that does not have much meaning in an information retrieval system. The reason that stop words must be removed from a text is that they make the text look large and holds less importance for analysts.

Stemming: Stemming techniques are used to find out the root or stem of a word. In the present work, the Porter Stemmer algorithm is used, which is the most commonly used algorithm in English.

Term Frequency—Inverse Document Frequency: Tf-idf is the product term for frequency and inverse document frequency. These are statistical methods. The number of times each term occurs in each document is counted and adds them all together.

Term Frequency—Term frequency (TF) is defined as the number of times a term occurs in a document

Inverse Document Frequency: Inverse document frequency is a method used for measuring the importance of a term in a text document.

(**b**) **Noun Phrase Identification**

To identify the noun phrases of a document, *tagging* should be done. Tagging assigns parts of speech to text and then noun phrases are identified. By extracting the combination of noun phrases, term frequency-inverse document frequency is calculated. And then algorithms or methods or techniques of research papers are identified.

4.2 Classification by Keywords

Categorization is identifying the main thesis of a document by placing the document into a predefined set of topics. Keywords are a set of important words in an article that give high grade description of its contents to readers.

4.3 Keyword Extraction

Keyword extraction is an important technique for various text mining-related tasks such as webpage retrieval, document clustering, document retrieval, and summarization. The main motive of keyword extraction is to extract the keywords with respect to their importance in the text.

Steps for keyword extraction are preprocessing, stop word removal, stemming, and tf-idf calculation which is mentioned in Sect. 4.1.2

Naive Bayes Algorithm for classifying area

Naive Bayes classifier is the most popular method for classification of text which helps in identification of whether a document belongs to one area(data mining) or classifying algorithms from module 1 into their category (clustering, classification, and association rule mining) and others by finding the word frequencies as the features. It depends on the precise or exact nature of the probability model; it can be trained in a systematic by a supervised learning.

Step 1: L : Set of feature words of order d_{xm},doc : input document to check
Step 2: Let there be ‘d’ Classes: C1, C2, C3...Cd
Step 3: Naïve Bayes classifier predicts X belongs to Class Ci if

$$\mathrm{P(Ci/X)} > \mathrm{P(Cj/X)}\ \text{for}\ 1 <= j <= m, j <> i$$

$$\mathrm{P(Ci/X)} = \mathrm{P(X/Ci)\,P(Ci)\,/\,P(X)}$$

Maximize P(X/Ci) P(Ci) as P(X) is constant

The probability of keywords occurrence will be calculated and the keywords with highest frequency count belonging to a particular area is identified.

4.4 Cue Word-Based Text Extraction for Demerits of Research Paper

Text Document D is taken as input

Cues word set Ci = $\{C1, C2, C3, C4...C_n\}$

Considering the conclusion part and

Segment the conclusion part into sentences Si = $\{S1, S2, S3, \ldots\ldots\ldots Sm\}$

Whenever any of the cue word in the cue word list (Ci) matches with words in the sentence list (Si) then extract the content of text document.

5 Results and Observation

The results obtained were from naturally randomly generated text documents and implementation of Java 7 on net bean Integrated Development environment. The following is the sample tagging result

```
Clustering/NN is/VBZ a/DT process/NN of/IN putting/VBG similar/JJ
data/NNS into/IN groups/NNS ./. Clustering/NN can/MD be/VB
```

```
considered/VBN the/DT most/RBS important/JJ unsupervised/JJ
learning/NN technique/NN so/RB as/IN every/DT other/JJ problem/NN
of/IN this/DT kind/NN;/: it/PRP deals/VBZ with/IN finding/VBG a/DT
structure/NN in/IN a/DT collection/NN of/IN unlabeled/JJ data/NNS ./.
This/DT paper/NN reviews/VBZ six/CD types/NNS of/IN clustering/NN
techniques-/FW k-Means/FW Clustering/NN,/, Hierarchical/JJ
Clustering/NN,/, DBSCAN/NN clustering/NN
```

Noun phrases are identified and key phrases are extracted from the document using *kea* algorithm, the list of key phrases up to (*n* = 4) are extracted. The results shown below are done in testing phase whereas 100 documents are considered and 80 for training phase and then 81–100 documents are given to testing phase. Below, four documents are shown and key phrases are identified for each document. The resulting key phrases are matched with the user defined dataset that consists of a list of algorithms and the obtained result for each document is shown below.

Document no	keywords	Keyphrase (*n* = 2)	Keyphrase (*n*= 3)	Keyphrase (*n* = 4)
81	27	10	3	–
82	13	2	–	2
83	22	–	1	–
84	18	3	–	–

Output can be seen below:

The algorithms extracted from the documents having the key phrase value *n* >= 2 which is mentioned below

Document no.	Extracted algorithms
81 [11]	k-means cluster, hierarchical cluster, DBSCAN cluster, bottom-up clustering method, top-down clustering method, partition methods, hierarchical methods, DENCLUE algorithm, density-based methods, grid-based methods, STastical Information Grid, Agglomerative algorithm, Divisive algorithm,
82 [12]	genetic algorithm, backpropagation algorithm, Naïve Bayesian learning algorithm, candidate phrase extraction algorithm
83 [13]	Transductive ranking algorithm
84 [14]	K-Means algorithm, DBSCAN algorithm, genetic algorithm

6 Conclusion and Future Work

This paper proposes extraction of algorithms from research papers and classifies its area and identifies the limitations of a paper. The advantage of this paper is reducing the search time for identification of algorithms, and classifies area or category of algorithms, and also gives drawbacks of paper within a short period of time. The proposed work is performed on a single domain (data mining) and the future work is to consider different domains.

References

1. Pacheri Bari, J.: Introduction of text mining and analysis of text mining techniques. Indian J. Res. **2**(2) (2013). Singhania University
2. Gupta, V., Lehal, G.S.: A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell. **1**(1), 60–76 (2009)
3. Sai Hari Priyanka, J.S.V., Sharmila Rani, J., Deepthi, K.S., Kranthi, T.: Information tracking from research papers using classification techniques. In: Satapathy, S.C. et al. (ed.) Emerging ICT for Bridging the Future, vol. 1, 153 Advances in Intelligent Systems and Computing 33. Springer, Switzerland (2015)
4. Wan, X., Xiao, J.: Single document keyphrase extraction using neighborhood knowledge. In: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)
5. Mooney, R.J., Nahm, U.Y.: Text mining with information extraction. In: Daelemans, W., du Plessis, T., Snyman, C., Teck, L. (eds.) Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, September 2003, pp. 141–160. Bloemfontein, South Africa, Van Schaik Pub., South Africa (2005)
6. Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., Wang, B.: Automatic keyword extraction from documents using conditional random fields. J. Comput. Inf. Syst. **4**(3), 1169–1180 (2008)
7. Menaka, S., Radha, N.: Text classification using keyword extraction technique. Int. J. Adv. Res. Comput. Sci. Eng. **3**(12) 2013
8. Joshi, A., Kaur, R.: Keyphrase Extraction in scientific articles: a supervised approach. Piaki Bhaskar Kishorjit Nongmeikapam Sivaji Bandyopadhyay **3**(3) 2013
9. Turney, P.: Learning to extract key phrases from text. National Research Council of Canada. Kim, S.N., Kan. M.Y. Technical report (2009)
10. Hanyurwimfura, D., Liao, B., Njogu, H., Ndatinya, E.: An automated cue word based text extraction. J. Converg. Inf. Technol. **7**(10) June 2012 doi:10.4156/jcit.vol7.issue10.50
11. Amini, M.-R., Usunier, N.: A Review: Comparative Study of Various Clustering Techniques in Data Mining. SIGIR’09, July 19.23, 2009, Boston, Massachusetts, USA ACM 978-1-60558-483-6/09/07
12. Sarkar, K., Nasipuri, M., Ghose, S.: A new approach to keyphrase extraction using neural networks. Int. J. Comput. Sci. Issues **7**(2), (2010)
13. Jadhav Bhushan, G., Warke Pushkar, U., Kuchekar Shivaji, P., Kadam Nikhil, V.: Incorporating prior knowledge into a transductive ranking algorithm for multi-document summarization. Int. J. Emerg. Technol. Adv. Eng. www.ijetae.com (ISSN 2250-2459, ISO 9001:2008), **4**(4) (2014)
14. Jadhav Bhushan, G., Warke Pushkar, U., Kuchekar Shivaji, P., Kadam Nikhil, V.: Searching research papers using clustering and textmining. Int. J. Emerg. Technol. Adv. Eng. **4**(4) (2014) www.ijetae.com (ISSN 2250-2459, ISO 9001:2008)