Original Research Article

J. Computing & Int. Systems (2022) 286–290

SACRED HEART RESEARCH PUBLICATIONS

Journal of Computing and Intelligent Systems

Journal homepage: www.shcpub.edu.in

ISSN: 2456-9496

A HYBRID APPROACH FOR INFORMATION EXTRACTION FROM TEXT

Ravi Lourdusamy [#1], Stanislaus Abharam[#2]

Received on 14 JUN 2022, Accepted on 29 JUL 2022

Abstract — A key term is a subset of terms or phrases from a text that can describe the meaning of the text. In our information age, key information terms are extremely useful for information retrieval, clustering, summarization, text mining, and text clustering, among other things. These are the terms from a text that can be used to describe the text's meaning. The primary goal of this paper is to assist users in quickly extracting key information from text using hybrid systems. The goal of the Hybrid system is to automatically extract important information from various texts. Linguistic and statistical approaches are used to extract key terms. These terms are then passed to a rule-based extractor for further refinement, where a statistical analysis is performed on this set of terms based on a variety of classes. Finally, this set is passed to Multi-layered Feed Forward Artificial Neural Networks, which extract key information terms via back propagation. Based on the performance evaluation, it was discovered that the acquired results are more efficient than manual judgement.

Keywords – Information Retrieval, Text Mining, Knowledge-Discovery in Text, Data Mining, Machine Intelligence

1. INTRODUCTION

Because of the fast growth of textual mat erial, IR is more vital than ever. The extraction of key information as an essential technique for document description has led to the focus on the field of information extraction. Key information phrases can assist users in quickly referring to documents to assess whether they are worth reading. As a result, essential information phrases are required. Automatic extraction of essential information terms requires no human intervention and speeds up the process of calculation to access and discoverability problems, bringing value to information organisation and retrieval. Because a key information term is the unit that expresses the meaning of an document, it can be used in a variety of applications such as automatic indexing, text summarization, IR, classification, clustering, filtering, cataloguing, topic detection and tracking, information visualisation, report gene ratio, web searches, and so on.

Key information phrases are crucial for IR, document retrieval, document clustering, summarization, Text Mining (TM), and text clustering, among other things. The group of terms from a document can explain the document's meaning. The primary goal is to extract crucial information from a document-utilising-hybrid structures that convey the entire meaning of the text.

Text mining has grown in importance as a research field. TM is the uncovering of previously undiscovered material by a computer by mechanically extracting information from various written resources. TM is a subfield of DM that seeks out intriguing patterns in vast collections. TM is a relatively new interdisciplinary field that combines IR, DM, MI, statistics, and computational linguistics.

The mechansim of extracting organised data from unstructured and/or semi-structured machine-readable publications is known as information extraction (IE). Automatic key IR is separating the small group of terms, key expressions, or key terms from a database that might reflect the significance of a document. Manually assigning high-quality keywords is costly, time-consuming, and error-prone. As a result, most methods and systems to assist people with automatic key term extraction have been presented.

Key information terms help readers determine whether or not a document is relevant to them. In the contemporary information age, key information terms are beneficial. As a result, these terms can be thought of as a set of terms or phrases that cover most of the text semantically. Key phrases provide a concise summary of a document's contents. As extensive document collections, such as digital libraries, become more common, the value of such summary information grows. Key terms and phrases are advantageous because they can be interpreted separately and independently of one another. They can be used in IR systems to describe the document returned by a query, as the foundation for search indexes, as a method of exploring a collection, and as a document clustering tool.

*** Corresponding author: E-mail:** [1]ravi@shctpt.edu
[2]astanislaus@gmail.com

[1] Research Supervisor , Department of Computer Science, Sacred Heart College (Autonomous), Tirupattur, Tamilnadu
[2] Research Scholar, Department of Computer Science, Sacred Heart College (Autonomous), Tamilnadu

2. Existing Techniques

Automatic key term indexing can be accomplished in two ways.

1) 2.1 Key term extraction

Terms that appear in documents are examined to find significant terms based on length and frequency. The goal here is to extract facts based on their significance in the text without prior vocabulary.

2) 2.2 Key term Assignment

Key terms are selected from a regulated vocabulary of terms, and documents are classified according to their content into classes that correspond to vocabulary components. Text Categorisation is the name given to this procedure. There is a prior arrangement of words, and the goal is to match them to messages in a collection. Some of the existing approaches for Automatic Key terms Extraction:

1. Simple Statistical Approach
2. Linguistic Approach
3. Methodology Based on Machine Learning
4. Graph-based Approaches

3. Proposed Technique

The proposed technique, Hybrid Information Extraction (HIE technique), is a hybrid technique. It combines statistics, linguistics, a rule-based algorithm, and a back-propagation algorithm. As illustrated in figure 1, it engages more than one methodology to solve a problem. The algorithms described are used for implementation, and research papers are used as data sets.

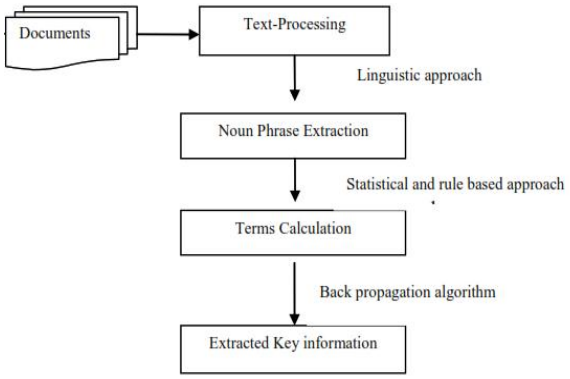

Figure 1: Hybrid Information Extraction technique

3) 3.1 Linguistic Approach

The method employs phonetics, such as highlighting terms, sentences, etc. It is more descriptive than prescriptive. A linguistic approach includes language analysis, which explicitly determines language's nature. Without the use of an algorithm, text processing can be done manually. Specific techniques and processes are used to assess current knowledge. It consists of syntactic analysis, discourse analysis, lexical analysis, etc. The linguistics approach used in this work is parts of speech tagging (POS), in which the terms in the entire article are tagged with parts of speech, and then only noun phrases are taken.

3.1.1 Text Pre-processing

Pre-processing a text is done in order to remove characters of the text that will not be useful for building ontology. The output of the text pre-processing is a list of essential words of the text that are required for further processing. In python, most of the pre-processing is done using libraries such as Regex and NLTK. Other libraries used for text pre-processing are Pandas, Unicode, Spacy and Gensim. Once the pre-processing of the text is complete, the list of words is sent for term extraction process.

3.1.2 Noun Phrases Extraction

Noun phrases are a combination of nouns, adjectives, or both nouns and adjectives. The steps taken in implementing noun phrases are as follows: first, the range of parameters for tagging nouns, adjectives, adverbs, and verbs are implemented. Subsequently, noun phrases are formed by combining nouns, adjectives, and both nouns and adjectives. The proposed algorithm is shown in Table 1 below.

Table 1: Steps for Linguistic Approach

Step 0. Start
Step 1: read the array list
Step 2: read the nouns list
Step 3: tag the terms which are nouns as NN
Step 4: read the adjectives list
Step 5: tag the terms which are adjectives as ADJ
Step 6: read the adverbs list
Step 7: tag the terms which are adverbs as ADV
Step 8: read the verbs list
Step 9: tag the terms which are verbs as VB.
Step 10: read the terms which are nouns, adjectives, and adjective + noun phrases
Step 11: Label the terms as NP.
Step 12: display the noun phrases that have been labelled as NP
Step 13: Stop

3.1.3 Statistical Approach

The statistical analysis aims to calculate the likelihood that differences as significant as or higher than those observed are due to chance. These methods are straightforward and do not require training data. The phrase statistics information can indicate significant information terms in the article. Some statistics methodologies are N-gram statistical information, term co-occurrences, TF*IDF, term frequency, and PAT-tree. The steps involved in the Statistical Methodology is depicted in Table 2.

Table 2: Steps for Statistical Approach

Step 0: Start
Step 1: read the array list
Step 2: Calculate the term's length.
Compute Length = Length (term)/ (Maximum_Length)
Step 3: Calculate the term's position
Evaluate term= #(term)/(∑ $term_j$)
Step 4: Calculate the term's phrase frequency
Compute F (frequency) = sqrt (0.5*pf*pf*plc)
Step 5: find the probability of t and a
Step 6: Check whether the term is present in the title or not
If present, t equals 1, otherwise t equals 0.
Step 7: Check whether the term is present in the abstract or not
If present, t equals 1, otherwise a equals 0.
Step 8: print the features list
Step 9: Stop

3.1.4 Rule-Based Approach

The Rule-based approach represents the learned model as a set of IF-THEN rules. Rules are an effective means of representing information or bits of knowledge. This study uses the rule-based technique to construct the knowledge base, and essential information terms have been split based on phrase frequency, abstract, and title aspects. The steps of Rule-based methodology are given in Table 3.

Table 3: Steps for Rule-Based Approach

Step 0: Start
Step 1: read the array list
Step 2: Determine whether the frequency of the term falls within the range
as per the user requirement
Step 3: find whether or not the term appears in both the title and the
abstract
Step 4: If this is the case, set the target value to 1.
Step 5: Otherwise, the target value is set to 0.
Step 6: Display the result according to the user requirement
Step 7: Stop

3.1.5 Back Propagation Approach

Back propagation is a neural network learning algorithm. A neural network is a collection of connected input / output units, each weight. The network learns throughout the learning phase by modifying the weights to anticipate the correct class label of the input tuples. Back propagation is a learning approach that uses a multi-layer feed-forward neural network to achieve learning. It generates a set of weights iteratively for predicting the class label of a tuple. A multi-layer feed-forward neural network comprises an input layer, one or more hidden layers, and an output layer. This work will use five input nodes, five hidden nodes, and a single output node. The knowledge base built through rule-based extraction and statistical analysis determines the model's target. This is implemented as follows: the weighted words with numerical weights derived from the features are fed into the feed-forward network, the learning process is carried out, and the error is computed. The goal, as mentioned earlier value is used to compute the error (either 1 or 0).The error rate is then compared to a defined threshold value. If the error rate is less than the threshold, the process is terminated and the output is given; otherwise, the weights are updated and the process is back propagated, and the error rate is calculated again, and this process is repeated iteratively until the error rate is less than the threshold value. This model's output is the key information words. The back-propagation algorithm is an n-n-1 network. It denotes that the input layer will have 'n' nodes, equivalent to the number of features used. The number of hidden nodes is equivalent to the number of input nodes, with a single output node indicating whether a specific term is a key information term or not.

Table 4: Steps for Back Propagation

Step 1: read the features list
Step 2: Initially, the specifications of the Back propagation method are accepted,
including hidden layer nodes, output layer nodes, learning rate, and angular
momentum.
Step 3: Weights for the network's links are accepted at random from -1 to 1.
Step 4: The model's epoch count is accepted.
Step 5: The dataset is segregated into training and testing sets using the hold
out method. Data set is randomly partitioned into two separate sets, a
training set and a testing set, in the hold out method. Typically, two-
thirds of the data is allotted to the training set, while the remaining
one-third is allocated to the test set. The training set is used to create
the model, whose accuracy is measured using the test set. The
estimation is pessimistic since the model is based on only a subset of
the initial data.
Step 6: The training data is delivered for training, and the weights are updated,
and these weights are used to calculate the output values of the testing data.
Step 7: Finally, the necessary key information terms are extracted.

4. Performance Analysis and Results

It is critical to assess the performance of the IR system that has been used as parts of the paper are described. A few conventional measurements, such as accuracy and error rate, should be used to evaluate the results. Accuracy is the state of being true; it is a correctness metric. The correct value of a standard is what is measured. The error rate is derived using the true and measured values. The error rate must be lowered as the number of papers increases. Similarly, as the number of papers increases, so does the accuracy. In the case of information extraction, these two elements are regarded as metrics for enhancing performance.

$$\% \text{ Error} = \frac{Truevalue - Measured\ value * 100}{Truevalue} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

% Accuracy= (100%) – (%Error) ……..(2)

Few documents are kept in the repository, and the key terms for these documents are extracted. In this section, the results of the manual assessment are contrasted to those of the system. For each research publication, a manual assessment is generated. A sample evaluation file for the first ten papers studied is presented below in Table 5.

Table 5: Comparison of manual and system generated

Document Number	Key terms generated manually	Key terms generated by the system
D1	6	5
D2	6	5
D3	4	4
D4	6	7
D5	8	6
D6	6	5
D7	5	6
D8	5	4
D9	6	5
D10	6	5

The values Accuracy and Error rate are determined as indicated in Table 6.

Table 6: Calculation of Accuracy and Error rate for articles

Number of Documents	Accuracy	Error
10	88.5%	13.5%
20	88.9%	13.3%
30	90.2%	12.4%

Table 6 demonstrates the system's performance for key term extraction across all types of documents.

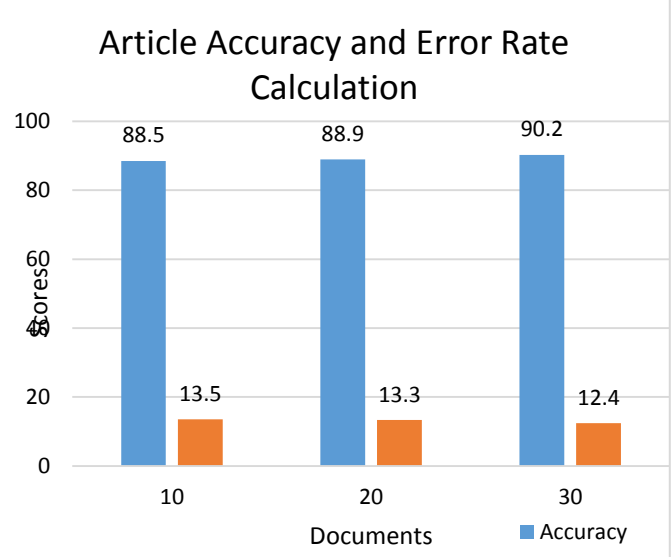

Figure 2: Accuracy and Error of the system

The system's accuracy and error are depicted in Figure 2. This demonstrates that the results gained are much superior. The x-axis shows the number of 'N' documents, while the y axis reflects the accuracy and error score. The technique produced about 89% accuracy.

5. Conclusion

The primary focus is on automatic key information extraction from documents utilising a hybrid system that employs various linguistic approaches, Statistical Approaches, Rule-Based Approaches, and back-propagation algorithms. It is feasible to avoid the restrictions and drawbacks of any different strategy by using the Hybrid approach. The user can extract vital information from the documents, such as key terms, using the proposed approach. The future work is a method of knowledge discovery using the key terms that are extracted from the document. The knowledge discovery process maps the concepts with a domain ontology by mining association rules for the concepts.

REFERENCES

1. Chengzhi Z ,Huilin W et al, "Automatic Keyword Extraction from Documents Using Conditional Random Fields", Journal of Computational Information Systems, Volume 4, issue 3, (2008).

2. Cohen J.D, "Highlights: Language and Domain- independent Automatic Indexing Terms for Abstracting" Journal of the American Society for Information Science, Volume 46 issue 3, pg no: 162-174, (1995).
3. Das A, Marko M et al, "Neural Net Model For Featured Word Extraction", Neural and Evolutionary Computing, ACM, (2002).
4. Damien Hanyurwimfura, Bo Liao et al, "An automated Cue Word based Text Extraction" Journal of Convergence Information Technology (JCIT), Volume7, Number10,(2012).
5. Ercan G, Cicekli I, "Using Lexical Chains for Keyword Extraction", Information Processing and Management, Volume 43 Issue: 6, pg no: 1705-1714, (2007).
6. Frank E, Paynter G.W, Witten I.H, "Domain-Specific Key phrase Extraction" Proceedings of the 16th International Joint Conference on Artificial Intelligence, Sweden, pg.no: 668-673, (1999).
7. Ion Muslea, "Extraction Patterns for Information Extraction Tasks: A Survey", AAAI Technical Report WS-99-11.
8. Jasmeen Kaur, Vishal Gupta, "Effective Approaches For Extraction of Keywords", IJCSI International Journal of Computer Science, Volume 7, Issue 6, (2010).
9. Kamal Sarkar, Mita Nasipuri and Suranjan Ghose, "A New Approach to Key phrase Extraction Using Neural Networks", IJCSI International Journal of Computer Science Issues, Volume 7, Issue 2 No 3, (2010).
10. Menaka S, Radha N, "Text Classification using Keyword Extraction Technique", International Journal of Advanced Research in Computer Science and Engineering, Volume 3, Issue 12, (2013).
11. Mihalcea R and Tarau P, "Text rank: Bringing order into texts", Association for computational linguistics, (2004).
12. Naidu Reddy et, al "Text summarization with automatic key word extraction in Telugu E-News Papers", (2017).
13. O. Medelyan, I. H Witten, "Thesaurus Based Automatic Key phrase Indexing", in Proceedings of the Joint Conference on Digital Libraries 2006, pg.no-296-297, Chapel Hill, NC, USA, (2006).
14. Parmar Paresh B and Ketan Patel "A Survey Paper on Mining Keywords Using Text Summarization Extraction System for Summary Generation over Multiple Documents" Volume 5 Issue 11, (2016).
15. Rahul B. Diwate, Prof. Satish J, Alaspurkar, "Study of Different Algorithms for Pattern Matching", International Journal of Computer Science (IJCSI) Volume 7, Issue 2, No 3, (2010).
16. Raymond J. Mooney and Un Yong , "Text Mining with Information Extraction" Proceedings of the 4th International MIDP Colloquium, (2003).
17. Yang, Shansong et. al "Key phrase DS: Automatic generation of survey by exploiting key phrase information", Volume 224, (2017).