Key Phrase Extraction System For Agricultural Documents

Swapna Johnny[1] and Dr. S. Jaya Nirmala[2]

[1] National Institute of Technology, Trichy, India
swapna.johnny@gmail.com
sjaya@nitt.edu

Abstract. Keywords play a vital role in extracting relevant and semantically related documents from a huge collection of various documents. Keywords represent the main topics covered in the document. But manual keyword extraction is a tedious and time-consuming process. Thus, there is a need for an automated keyword extraction system for easier extraction of relevant documents. In this paper, the focus is on agriculture-related documents. Agrovoc, an agriculture-based vocabulary that contains more than 35,000 concepts is used for extracting relevant keywords from the document. The proposed system extracts the relevant keywords from agricultural documents which are further used to extract relevant documents. Also, with the increasing number of documents on the Internet, the need for efficient storage for keywords with their corresponding documents is necessary. In the proposed system, a trie-based inverted index has been used for efficient storage and retrieval of keywords and the related documents.

Keywords: Keyword Extraction, AGROVOC, Agrotags, Agricultural Documents, Trie-Based Inverted Index.

1 Introduction

Farmers are a vital part of our society. Their queries and doubts related to agriculture need to be answered to increase crop production. There are many agricultural thesauri available, but the extraction of keywords is a manual process, thus it makes it difficult to analyze the document. Also, it is not possible for the farmers to go through the entire document. Thus, an automated query answering system based on the agriculture thesaurus is proposed. The first phase of the project deals with the extraction and storage of the keywords from various agricultural documents. The task of identifying the set of the terms that characterize a document is essential for information extraction. These terms that signify the most important information in the documents are called by various names: *key terms*, *key phrases* or just *keywords*. All these words have the same purpose – to represent the important concepts discussed in the document. With the increasing number of digital documents present on the Internet, key phrases are essential and make important metadata. Many documents still don't have keywords assigned to them and

the process of manual assignment of keywords largely depends on the subject experts and authors. Thus, the need for an efficient keyword extraction is vital.

With agricultural documents, there are many technical terminologies that may not be commonly used and are listed only in agricultural thesaurus like AGROVOC, Centre for Agriculture and Bioscience International (CABI), National Agricultural Library (NAL) etc. Documents like these prove to be crucial for keyword extraction from various agricultural documents and are extremely useful in agriculture information systems. Linking semantically related agriculture documents can be made easier by analyzing and extracting keywords from these documents. With the use of natural language processing techniques and AGOVOC, extraction of key phrases can be done and thus the processing of these documents is made more efficient and faster.

Another issue related to information extraction is the storage of the extracted keywords and the retrieval of the corresponding documents. The searching and retrieval time should be minimized. An efficient way of doing the same is by using an inverted index. Almost all retrieval engines for full-text search today rely on an inverted index, which gives access to a list of documents related to a term when the term is given to the inverted index. An even more efficient indexing method can be achieved by combing two data structures – inverted index and trie. An inverted index helps in efficient indexing and trie helps in faster retrieval time.

The agricultural documents used in the system are taken from AGRIS (International System for Agricultural Science and Technology) which consists of more than 8 million structured bibliographical records on agricultural science and technology. The agricultural thesauri AGROVOC is an RDF file consisting of more than 35,000 concepts in multiple languages in the form of triples. It is made available by the Food and Agriculture Organization of the United Nations (FAO).

2 Keyword extraction approaches

2.1 Supervised Approach

This approach refers to the system learning to predict keywords from an already tagged document set. The keyword extraction system is trained with the training document set and then used to predict keywords from new documents. One example of this approach is the Keyphrase Extraction Algorithm (KEA). On the training documents, using TF-IDF and first occurrence, KEA applies Bayes theorem to extract the key phrases from new documents [6]. In KEA, TF-IDF stands for Term Frequency Inverse Document frequency. A word having a higher TF-IDF implies that the term occurs frequently only in that document and not in any other document. Thus, the word is a keyword.

2.2 Unsupervised Approach

This approach refers to the system learning to predict keywords without any training document set. The system learns by itself on how to extract keywords. Two examples under this approach are TextRank and Rapid Automatic Keyword Extraction (RAKE).

TextRank extracts key phrases by creating a graph from the text and ranking the key phrases based on the co-occurrence links between words. The advantage of TextRank is that it is language independent as its extraction process depends on the word occurrences [11]. RAKE is also a language-independent keyword extraction process. It works on the assumption that keywords do not contain stop words and punctuation.

3 Existing Work

Agrotagger, an automated key phrase extractor for various agricultural documents, identifies the agrovoc terms from the document and maps them to Agrotags terms. The paper [1] describes a mapping process which consists of narrowing down the keywords. The agrovoc terms are mapped according to the broader term and non-descriptor term. All the non-descriptor terms get mapped to a single Agrotag. For example, 'paddy' the non-descriptor for 'rice' is mapped to 'rice'. Then various terms are again mapped down to the broader concept like, for example, 'garden waste' is mapped to 'organic waste'. Also, during the mapping process, all the non-descriptors, scientific names, geographical terms, and fishery-related terms are removed [1]. Agrotagger helps to link documents related to agriculture so that there is a more efficient and faster retrieval of information.

The paper [6] deals with the study and use of social networking sites for agro-produce marketing. It helps to connect the farmers directly with the merchants. The posts on various social networking sites can be analysed to retrieve relevant information which can be propagated to the farmers and merchants as suggestions. The information extraction system extracts entities and related data by using domain-specific named entity recognizer from the tweets on Twitter. This extracted information is then exchanged between the farmers and merchants [7].

The paper [12] deals with the process of unsupervised keyword extraction by using noun clustering. In the system proposed in the paper, two extracted nouns are in the same cluster if they have at least one common word. Then keyphrase extraction is done by choosing the top k clusters, scores of which are calculated by averaging the scores of all the noun phrases in that cluster. The score of each noun phrase is the sum of the frequencies of each individual word in that phrase. The paper has shown that keyword extraction is more efficient if nouns are clubbed with their synonyms [12].

An inverted index is a data structure that contains a list of all the distinct words that are present in all the documents, and for each word, a list of the documents in which it appears is also present. The inverted index helps to map the words to its corresponding documents.

An inverted index consists of two parts:

- Vocabulary – set of all unique words
- Occurrences – list containing all the important information related to the word like frequency of word, documents in which the word appears

Trie is an information retrieval tree data structure used to store multiple strings in which each node has a maximum of 26 children, one for each alphabet. The last node of a string is denoted by *EndOfWord* field in the node. All the descendants a node have

a common prefix of the string associated with that node. Fig. 1. shows an example of example of a trie data structure.

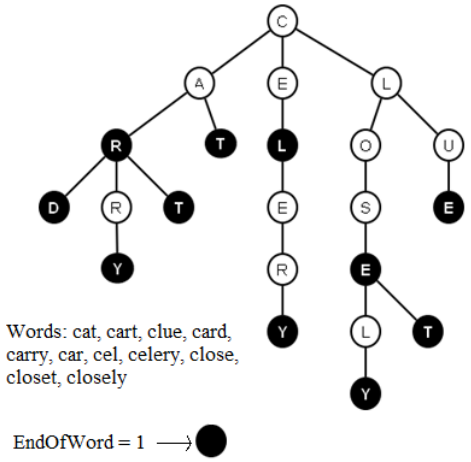

Fig. 1. Trie Data Structure

The advantage of trie is that the time complexity of searching is reduced to *O(m)* where *m* is the maximum length of the string. Thus, searching and information retrieval is very efficient with this data structure. It can be used to store an inverted index. The index which contains the list of documents in which the keyword is present can be placed alongside *EndOfWord* field. The paper [4] describes the use of a trie as an inverted index for a search engine. The performance of a search engine depends on the document searching and retrieval time and thus, the proposed data structure facilitates the fast searching of relevant information over the web. The system removes the step words, stems the keywords and then creates the index. Thus, the user queries are processed much faster with the proposed system [4].

4 Proposed solution

The proposed solution makes use of natural language processing techniques for the extraction of keywords and a trie data structure for the indexing and storage of the thus obtained keywords. A text file consisting of links to different agricultural documents present on AGRIS (International System for Agricultural Science and Technology) is given as input by the user for the keyword extraction process. Once the keyword extraction is done with the help of AGROVOC, the extracted keywords and its corresponding documents are indexed by using a trie data structure. Fig. 2. shows the block diagram of the proposed solution.

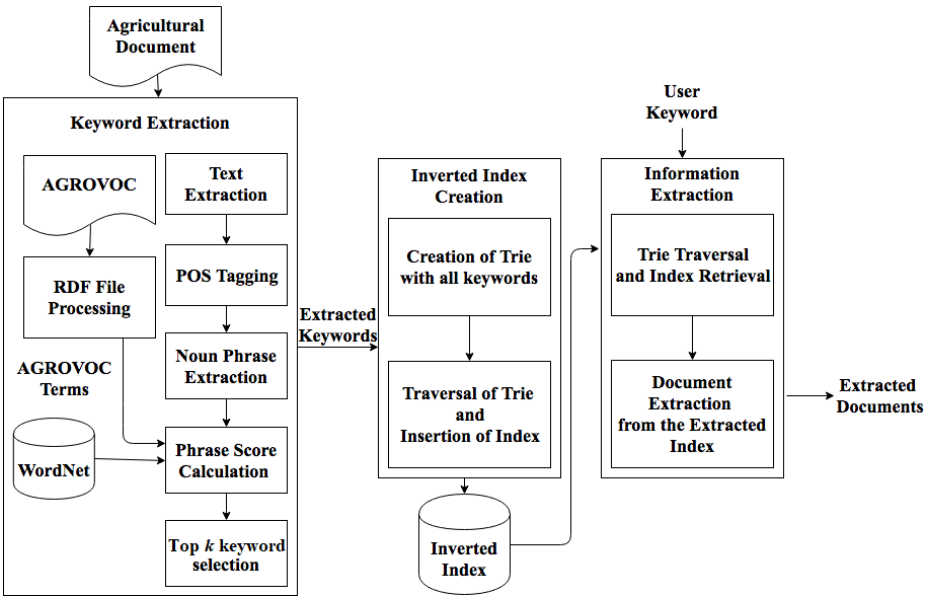

Fig. 2. System Block Diagram

4.1 Keyword Extraction

Keyword extraction is an important process to identify the major concepts and terms present in the document. It is easier to select which document to read if the appropriate keywords are extracted. Here, an unsupervised keyword extraction technique is proposed where different natural language processing techniques and statistical techniques have been used for the extraction process. In this module, for each of the agricultural documents, the corresponding keywords are extracted with the help of AGROVOC and WordNet. These extracted keywords help to identify what the document is about and are thus essential for information extraction.

RDF File Processing AGROVOC contains all the agriculture-related terms and concepts. It is an RDF (Resource Description Framework) file and processing is required to extract the AGROVOC terms from it. The RDF file is used to represent information or resources present on the web and is made up of triples each of which consists of 3 components – subject, predicate and object. These three components are also known as resource, property and property value. The AGOVOC RDF file used in this work consists of triples that describe schemas and concepts. To access the AGOVOC terms, only those triples corresponding to a concept are required. An example of a triple is given below:

```
<rdf:Description
```

```
rdf:about="http://aims.fao.org/aos/agrovoc/c_30790">
 <rdf:type rdf:re-
source="http://www.w3.org/2004/02/skos/core#Concept"/>
</rdf:Description>
```

In the triple, <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/> indicates that the triple is that of a concept. <rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/c_30790"> gives the link to the AGROVOC term.

SPARQL, a query processing language for RDF files, is used for processing each triple and selecting those triples which correspond to a concept and extracts the link to the AGROVOC term. Since AGROVOC consists of agricultural terms in multiple languages, the English name of the concept is extracted.

Text Extraction. Since the input is a text file consisting of links to agricultural documents taken from AGRIS, the system needs to identify and then extract the text from its corresponding web page. The extracted text is then divided into sentences and these sentences are then further processed by the next block.

Parts of Speech Tagging. In a sentence, each word belongs to a category under parts of speech (POS) like noun, adjective, adverb, verb etc. In this module, the POS of each word in the sentence is assigned. This step is important because the same word can have a different POS tag in different contexts. Thus, word sense disambiguation is done here for identifying the meaning of the word when the same word may have different meanings. Each word along with its POS tag is forwarded to the next block for further processing.

Noun Phrase Extraction. Based on the POS tagging of each word, the noun phrases are extracted in this block. This is essential to the keyword extraction process because the noun represents most of the important information in a document and thus extracting noun phrases would simplify the keyword extraction process. The various possible patterns for a noun phrase are given below:

- Noun
- Nouns
- Adjectives Noun

To identify a noun phrase, a regular expression containing the different patterns for a noun phrase is used to extract the noun phrases. The regular expression used by the system is <NN.*|JJ>*<NN.*> where NN represents a noun and JJ represents an adjective. The system uses the regular expression and looks for the matching pattern in the sentence with the POS tags. The identified noun phrases are then forwarded to the next block for further processing.

Phrase Score Calculation. In this block, each phrase is given a score based on the frequency of similar words. WordNet, a database consisting of words and its synonyms, helps to calculate the similarity between two words. Higher the value more similar the words are. It's been shown that the co-occurrence distribution of words in different phrases represents how important a phrase is in a document. A phrase *w* is a key phrase if the probability distribution of the phrase *w* and the frequent phrases is biased to a subset of frequent words [5]. Thus, in order to calculate the phrase score of a phrase, the frequency of each word in the phrase is calculated.

Given phrase = word[1] word[2] … word[n]

$$\text{Phrase\_Score} = \sum_{i=1}^{n} (frequency[word[i]]) \quad (1)$$

where n = number of words in the phrase

Top *k* keyword selection. The phrases are sorted according to their phrase scores and from these phrases, the top *k* phrases are selected as the extracted keywords.

4.2 Inverted Index Creation

An inverted index is a data structure that is used to map the words to the documents it is present in. One of the major disadvantages of using a traditional inverted index is that there is a large storage overhead and the cost of insertion, updation, and deletion increases with the increase in the number of keywords. Trie data structure's advantage is that it can store many words in limited space and the search time is also drastically decreased to the length of the keyword. Thus, in the proposed system a trie-based inverted index is used. It combines the advantages of both an inverted index and trie for efficient and fast information retrieval system.

Trie Node creation. For every keyword extracted, first, the trie is traversed to check if the keyword already exists in the inverted index. If the keyword doesn't exist, the trie is traversed until the longest matching prefix is found and then for the remaining part of the word the corresponding nodes are inserted into the trie.

Insertion of Index. Once the keyword is found in the trie, the document number to which the keyword corresponds to is appended to the list of documents in the leaf node of the keyword.

4.3 Information Extraction

Retrieval of documents related to the user entered keyword is done in this block. The trie based inverted index is traversed to locate the document numbers in which the keyword is present.

Trie Traversal and Index Retrieval. The trie based inverted index which stores all the keywords and its corresponding documents is traversed to locate the user entered keyword. The index consisting of the list of related documents is obtained in the last node of the user entered keyword in the trie. That index is then forwarded to the next block.

Document Extraction from the Extracted Index. In this module, the documents in the index are extracted and displayed to the user. These documents that are extracted are the documents that are related to the user entered keyword.

5 Results

5.1 Extraction of AGROVOC Terms

AGROVOC contains more than 35,000 concepts related to agriculture, food, nutrition, fisheries, forestry, environment etc. The AGROVOC RDF file needs to be processed in order to obtain the AGROVOC terms. A total of 35859 terms were extracted by the system. Some of the AGROVOC terms extracted after processing the AGROVOC RDF file are shown in Fig. 3.

```
swapna@swapna-Lenovo-ideapad-500-15ISK:~/Documents/NITT/Project$ python rdf.py
Total no of concepts =  35859
1   Cottocomephorus grewingkii
2   Scomberesox saurus
3   neurotoxicity
4   Edovum puttleri
5   sustainable development
6   Clarias werneri
7
7   random sampling
8   Zygophyllaceae
9   radiation
10
10   international relations
11   Thysaniezia
12   Acremonium coenophialum
13
13   Metcalfa pruinosa
14   Oxymonacanthus halli
15   Parthenium
16   Chagas' disease
17   isoelectric focusing
18   Mullus
19   Scabiosa caucasica
20   Phycis
```

Fig. 3. AGROVOC Terms

5.2 Extracted Keywords

After applying Natural Language processing techniques, the top *k* phrases with maximum phrase score are selected as the extracted keywords. The keywords extracted for Document 1 is given below:

Keywords for Document 1
['PPR VIRUS', 'PPR', 'INDIA', 'SHEEP', 'TOTAL', 'LYMPH NODE', 'DISEASE', 'BLOOD', 'RUMINANTS', 'RINDERPEST']

5.3 Extraction of Documents from user entered keyword

The system reads the user entered keyword and then the documents which contain the user entered keyword are displayed. The documents in which the keyword “RICE” exists is displayed in Fig. 4.

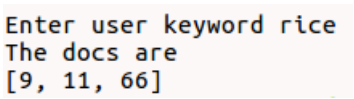

Fig. 4. Extracted Documents from User Entered Keyword

5.4 Experiment evaluation

Table 1. denotes the evaluation metrics for the system.

Table 1. Evaluation Metric

Metric	Value
Accuracy	91.79%
True Positive Rate (Sensitivity)	82.75%
Precision	63.06%

For evaluating the performance of the system accuracy, true positive rate and precision metrics have been used. The accuracy calculated was 91.79%, true positive rate was 82.75% and precision was 63.06%.

Since the number of keywords to be predicted is lower than the words that are not supposed to be predicted, the true positive value is always lower than the true negative value. Thus, even if the correctly predicted keywords are less, the effect on the accuracy is less because the true negative is large. This is the reason why the true positive rate is lower than the accuracy rate. The accuracy obtained for each of the documents is depicted in Fig. 5. The overall average accuracy obtained is 91.79%.

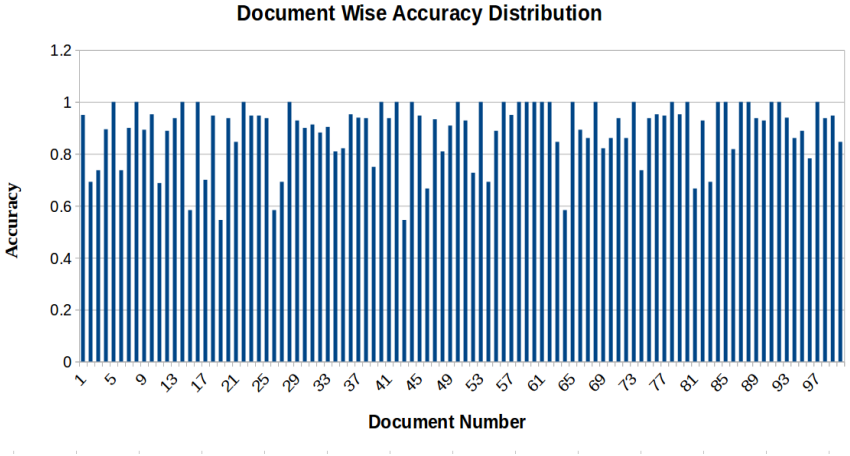

Fig. 5. Document Accuracy

Although the overall accuracy was calculated as 91.79%, the range of accuracy is from 56% to 100%. Analysis of the system denotes that one of the reasons of lower accuracy in some documents is that the system is unable to predict chemical compounds like Nitrogen (N), Phosphorous (P) etc. since the keyword is the chemical name, but the chemical symbol is mentioned in the document. Another reason is that the keywords that are not noun phrases are not detected. Keywords that are verbs will not be detected as keywords by the system. This again lowers the accuracy rate. The value of *k* in the top *k* keyword selection also affects the accuracy rate. A lower value of *k* may lead to the elimination of the keyword due to a lower phrase score. A higher value of *k* may lead to the inclusion of all keywords but will also detect words that are not keywords.

True positive rate (sensitivity) denotes the number of correctly predicted keywords out of the actual keywords. Since the number of actual keywords is usually around 5-7 keywords, a small decrease in the number of predicted keywords causes the true positive rate to decrease drastically. Also, with a larger value of *k* in the top *k* keyword selection, the precision increases because a larger set of keywords is extracted.

Precision denotes the number of correctly predicted keywords out of all the keywords predicted by the system. Since the value of *k* taken by the system is usually larger than the actual number of keywords, the precision is low. This implies that a superset of keywords is extracted. As the value of *k* increases, the precision decreases.

Table 2. Evaluation Metric Values for different *k* values

Metric	***k* = 5**	***k* = 10**	***k* = 15**	***k* = 20**
Accuracy	94.59%	91.79%	81.73%	71.63%
True Positive Rate (Sensitivity)	70.32%	82.75%	88.96%	91.26%
Precision	70.71%	63.06%	52.93%	48.46%

The value of all the evaluation metrics used is directly proportional to the number of correctly predicted keywords. The changes in the evaluation metric values for different *k* values are shown in Table 2. One of the reasons why the correctly predicted keywords are less is that the frequency of the words in the key phrase is less in the document which causes its phrase score value to decrease. For example, the keyword 'nutrient availability' occurs in a document only once and then the document describes the effects of various substances on the different nutrients such as potassium and phosphorous available in the soil. A lower phrase score value compared to the other phrases means that the phrase is not selected as the key phrase. Since the number of keywords correctly predicted decreases, the true positive rate and precision decreases. Accuracy also decreases but not as much as the other two metrics.

Another reason for the keywords to be not predicted is that the phrases extracted are only noun phrases. If the keyword to be predicted is a verb then the keyword will not be detected. Thus, there is a decrease in the number of correctly predicted keywords. Another reason for incorrect keyword prediction is that the document contains the abbreviation or chemical composition, but the keyword is the full form or the chemical name.

5.5 COMPARISION

The system is compared with Summa 1.1.0 which is the implementation of TextRank in python 3. TextRank extracts keywords by creating a graph from the text and ranking the key phrases based on the co-occurrence links between words. The comparison between the two techniques is given in Table 3.

Table 3. Comparison between the proposed system and TextRank

Metric	Proposed System	TextRank
Accuracy	91.79%	67.88%
True Positive Rate (Sensitivity)	82.75%	87.28%
Precision	63.06%	44.06%

The number of keywords extracted by TextRank is comparatively higher than that by the proposed system. TextRank extracts a superset of the keywords. One of the reasons behind the large set of keywords is that two similar terms are treated as separate keywords. WordNet does not contain the synonyms for many phrases. The system can handle the cases where the number of words in two similar phrases are the same and each corresponding word is similar. For example, 'health hazard' and 'health risk' are similar phrases but WordNet does not contain these phrases. The system thus compares the corresponding words, 'health' with 'health' and 'hazard' with 'risk'. In such a case, the system can treat both the phrases as similar. TextRank on the other hand treats 'health hazard' and 'health risk' as two different keywords. Because of the larger keyword set, the precision of TextRank is lower compared to that of the proposed system. But because of the same reason, the True positive rate is higher for TextRank in comparison with the proposed system. Since the extracted keywords are more in the case of

TextRank, it also implies that there is an increase in the number of false positives and thus a decrease in the number of true negatives. Thus, the overall accuracy for TextRank is lower in comparison with the proposed system.

6 Conclusion

This paper examined a system for extraction and indexing of keywords from agricultural documents with the help of AGROVOC. The system uses various natural language processing techniques to extract keywords from the documents and makes use of a trie-based inverted index data structure for indexing. The user entered keyword is searched in the inverted index and the corresponding documents are extracted. It was found that the accuracy of the prediction of the keywords was 91.79%, the true positive rate was 82.75% and the precision was 63.06%. The performance of the system for different values of *k* was performed and analyzed. The proposed system was also compared with TextRank and the results were analyzed. The system, however, was not able to detect keywords that were verbs and whose chemical composition was mentioned in the document.

References

1. Balaji V. et al.: Agrotags – A Tagging Scheme for Agricultural Digital Objects. In Metadata and Semantic Research Communications in Computer and Information Science. vol. 108, pp. 36-45. Springer Berlin Heidelberg, (2010).
2. Cutting, Douglas & O. Pedersen.: Optimizations for Dynamic Inverted Index Maintenance. In Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval pp 405-411, Jan. (1990)
3. H. F. Sun and W. Hou.: Study on the Improvement of TFIDF Algorithm in Data Mining. Advanced Materials Research, Vol. 1042, pp. 106-109, 2014.
4. Priyanka S Zaware and Satish R Todmal.: Inverted Indexing Mechanism for Search Engine. In International Journal of Computer Applications, 123, 15-19, (2015).
5. Matsuo, Yutaka & Ishizuka, Mitsuru.: Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. In International Journal on Artificial Intelligence Tools, 13, pp. 157-169, (2003).
6. Siddiqi, Sifatullah & Sharan, Aditi.: Keyword and Keyphrase Extraction Techniques: A Literature Review. International Journal of Computer Applications. 109. 18-23, (2015).
7. P. Joshi, S. Chaudhary and V. Kumar.: Information Extraction from Social Network for Agro-produce Marketing. In International Conference on Communication Systems and Network Technologies, Rajkot, pp. 941-944, (2012).
8. Luthra, Shruti & Arora, Dinkar & Mittal, Kanika & Chhabra, Anusha, A Statistical Approach of Keyword Extraction for Efficient Retrieval. International Journal of Computer Applications. 168. 31-36, (2017).
9. Terrovitis, Manolis & Passas, Spyros & Vassiliadis, Panos & Sellis, Timos.: A combination of trie-trees and inverted files for the indexing of set-valued attributes. Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 728-737, (2006).

10. B. Balcerzak, W. Jaworski and A. Wierzbicki.: Application of TextRank Algorithm for Credibility Assessment. In IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, pp. 451-454, (2014).
11. Shuo-Fu Yen, Jiann-Jone Chen, Yao-Hong Tsai.: Efficient Cloud Image Retrieval System Using Weighted-Inverted Index and Database Filtering Algorithms. Journal of Electronic Science and Technology, 15(2), 161-168, (2017).
12. M. Rezaei, N. Gali and P. Fränti, ClRank.: A Method for Keyword Extraction from Web Pages Using Clustering and Distribution of Nouns. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Singapore, 2015, pp. 79-84, (2015)
13. AIMS AGROVOC HomePage http://aims.fao.org/vest-registry/vocabularies/agrovoc, last accessed 2019/01/30.