ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Keyword Extraction Using Swarm Intelligence Techniques

Sheeba.J.I, Sowmya.D, Pradeep Devaneyan.S

Assistant Professor, Dept. of CSE., Pondicherry Engineering College, Pondicherry, India

PG Student, Dept. of CSE., Pondicherry Engineering College, Pondicherry, India

Professor, Dept. of Mechanical Engineering, Christ College of Engineering and Technology, Pondicherry, India

ABSTRACT: Without formal structure data are those that have no prearranged form or structure and are full of textual data. Typical unstructured systems include emails, reports, telephone or messaging conversations, etc. Keywords are frequently used in multimedia resources and document databases to establish information and regulate if two pieces of test are relevant to each other. Extracting keywords is major complicated problem when working with text. The main goal of this work is to extract the keywords from text conversation using Swarm Intelligence (SI) optimization techniques. Keywords are grouped together under their classification and then suggested to the user. By using crisscross searching techniques, keywords are easy to search and provide some important top keywords. Separate algorithms are used for searching and clustering. Firefly algorithm is used for clustering to identify similar related keywords. Finally the results show that SI techniques have better effect and convergence performance compared to existing diverse keyword extraction methods.

KEYWORDS: Swarm Intelligence, Optimization, Crisscross, Firefly, Searching, Clustering.

I. INTRODUCTION

The lot of information accessible by Internet is very large. It is ambitious over complete text material. So there is a precondition of good keyword extraction and searching methods this can provide contents of a given text in particular manner. As keywords are smallest unit of information, they can provide a compact portrayal of a document's content. Many existing approaches for keyword extraction build on human indexers for standard assignment of Keywords. Manual keyword extraction uses rigid taxonomy, is time exhausting and a difficult task [1]. Research is accordingly required to focus on methods that can extract keywords from text document. This can act as a benefit to extract keywords in documents that would otherwise be inaccessible. There are many ways by which keyword extraction can be carried out, such as statistical methods, supervised and unsupervised machine learning, diverse extraction methods and linguistic ones. Statistical methods for the extraction of keywords from documents have some advantages by linguistic based approaches alike as; the same approach can be enforced to many different languages beyond the obligation to develop different set of rules any time for a different language. With the increase in time of enormous number of textual data in different languages, it is now available to adequately apply different statistical approaches for keyword extraction and accomplish good results [3].

A few works has been concluded for extracting keywords using statistical approaches; frequency of a word is an index of the influence of that word in the document. Altogether more considerable a word is to a document, the more number of times it should appear in the document. Considering in any document the words having highest frequencies are normally the stop word words (the, but, of, is etc.), one solution is to privileged a frequency threshold raised which most the words are stop words and below which are keywords [4]. However, such a threshold doesn't endure in practice. Other approaches have been proposed such as Term Frequency (TF) and Table Term Frequency (TTF) which work on a collection of documents. TTF developed the score of a word in proportion to its frequency and decreases it in proportion to the static documents in which it emerge in the corpus. TF score gives much better results than frequency score but it can't be used for extracting keywords from a single document [5].

DOI: 10.15680/IJIRCCE.2016. 0404066

Likewise, it is relevant to develop approaches which can be used on capability of poor languages (non-availability of corpus or finite number of text documents etc). In existing, Diverse and persistent lists of keywords, which can be recommended to the user of a text document to achieve their information needs without, distract them. These lists bring back frequently by submitting multiple implicit queries derived from the distinct words. Each query is associated to one of the topics analysed in the conversation preceding the recommendation, and is acknowledged to a search engine over the Wikipedia. The topic positioned clustering decreases and the diversity of keywords increases the opportunity that slightly one of the recommended keyword answers a need for information, or can advance to a useful to the user [6].

In this framework, it includes identifying the keywords, feature extraction, using swarm intelligence optimization techniques are used to reduce the words and grouping the keywords depending upon the user choice.

II. RELATED WORKS

In keyword Extraction method (T.Jo, 2016) proposed Table based KNN for extracting keywords. The main task is word classification from encoding words into table and provides a better keyword extraction result for each text [4] .In (M.Habibi and A.Popescu-Belis 2013) proposed diverse keyword extraction from conversation using diverse keyword method. To extract the diversity of topics and then influenced from summarization, this method exaggerate the coverage of topics that are identified automatically in transcripts of conversation fragments [2]. In (Jianxin Li, Chengfei Liu 2015) they introduced context based diversification for keyword queries over XML data. It fundamentally diversifies XML keyword search based on its different contexts in the XML data by inclined a short and vague keyword query and XML data to be searched, to derive keyword search candidates of the query by a simple feature selection model[3]. In (M.Habibi and A. Popescu-Belis 2014) they proposed enforcing topic diversity in a document recommendation for conversation. In addition of extracting keywords Significant and relevant lists of documents, which can be recommended to the participants of a conversation to accomplish their information needs without entertain them [1]. In (Kyoungrok Jang, Kangwook Lee , 2016) Food hazard event extraction based on news and social media:A preliminary work. The main aims to detect and extract food hazard event from the live data shared on the Web[]. In (Vishal Gupta, Gurpreet Singh Lehal) proposed a survey of text summarization techniques based on text extraction techniques mainly uses linguistic methods to inspect and illustrate the text and then to asset the new concepts and expressions best to characterize it by making a new shorter text that transfer the uttermost extensive information from the original text document [10]. In (P.Wu, X.Zhang, 2016) proposed a Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion. Keyword spotting strategy based on decision fusion is proposed in order to make the best use of audio-visual speech and adapt to diverse noise conditions. Weights generated using a neural network combine acoustic and visual contributions [11].

In most existing approaches are based on text document classification. In expansion of this proposed work it is going to cluster the keywords from text using diverse keyword extraction [6]. In this proposed model it is going to cluster the keywords by using swarm intelligence optimization techniques and also recommended to the user. Here it also introduces some additional features for improving keyword extraction methods like term frequency, table term frequency and SI techniques using reducing keywords. It will improve the quality of the keywords.

The remaining of this paper is coordinated as follows. Section 3 presents proposed framework for keyword Extraction, Section 4 for experimental results and Section 5 conclude the paper.

III.PROPOSED FRAMEWORK

The below framework is proposed to extract keywords from the given input using swarm intelligence techniques. The processing of keyword extraction involves two major problems, first to extract the keywords and second to cluster the keywords. Fig 1 shows block diagram for keyword extraction. First step is to select the desired input text and it can be preprocessed. In preprocessing, stemming and stop word removals are applied in order to reduce noise and conflicting attributes. Feature extraction is used to obtain the important words in the text. After preprocessing, each word from the input is represented by a vector of features. Here term frequency is calculated and to find out the total number of words occurs in a keyword list. The calculated keywords are said to be table term frequency. Because the terms possessed are stored in a temporary table and intent some threshold value. Here extract all keywords in the list. Using crisscross search algorithm searching the keywords is easy and it provides nearest user searched keywords. It will show some global best keyword list and it consider as a top N keywords. Then user selects one of the keyword in that list and firefly clustering algorithm is used for clustering. Finally Clustered keywords result is to be displayed according to their user choice.

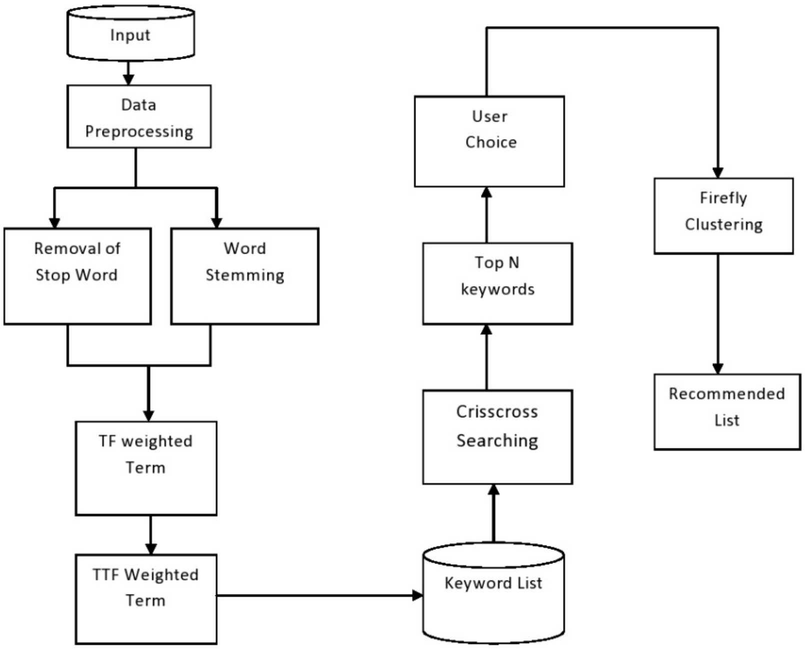

Fig 1: Proposed Framework for Extracting keywords and clustering from conversation

The following steps are maintained in this proposed framework:

Data Preprocessing:

In data preprocessing is used for regenerating some preprocessing tasks is commonly achieved. Data preprocessing includes Feature extraction, term frequency calculation and table term frequency .This paper has attempted stemming and stop word removal process for data preprocessing [3].

Stemming is the process for contracting the inflected words in prefix or in suffix. The Stemming process makes the word shorter by removing such things as plurals and gerunds. A stem is a form to which affixes can be attached. Examples for stemming: The words attaching, attaches, attached can be changed into attach after the stemming process.

Stop words are elimination of the words which appear regularly in the input giving lesser meaning while identifying the fundamental content of the input. There is not one explicit list of stop words based on human input. To save memory space and speed up search results in order to remove the stop words here. Examples of Stop words A, The, As, Of etc.

Feature Extraction:

The Feature extraction techniques are used to ingress the important words in the text. After preprocessing, each word from the input is defined by a vector of features. For each word, successive two features are enforced such as term frequency calculation [2].

Frequency Calculation: In the first feature, it is calculated by counting the number of each word occurring in the preprocessed data. The frequency is calculated by the occurrence of the word in which the highest frequency is calculated.

Table Term Frequency: In this calculate a threshold value. Then collect terms, whose weight is above the threshold value. Here, the threshold is the most important terms from each keyword according to TF value. Then count term frequencies from the term collected. To call this term frequency as “Table Term Frequency” because the terms collected are stored in a temporary table.

Crisscross Searching:

After feature extraction, crisscross algorithm is to be used; it is used to extract the maximum possible pbest keywords. Here extraction of keywords from conversation for which keyword list must be recommended, as provided by the system. These keywords should cover as much as possible the topics detected in the conversation. The crisscross optimization algorithm (CSO) is a recent evolutionary algorithm mainly used for searching. Related to other heuristic algorithms, the CSO algorithm has a enormous advantage in solution accuracy and convergence speed when addressing complex optimization problems [4]. The CSO algorithm is primarily made of three components: the horizontal crossover, the competitive operator and the vertical crossover. Both horizontal crossover and vertical crossover execute in individual iteration and replicate their offspring result called moderation solutions by achieving different crossover operations. Simultaneously, later each operation, the competitive operator must be carried out to prefer the better particle. Only those new solutions that exceed their parent particles can recover while the others will be eradicated in the competition. It is the sequence of the double search strategies and competitive mechanism that implement CSO important advantages in convergence speed and solution accuracy [6].

Steps for Crisscross Searching Algorithm:

Step 1: Define input Data
Step 2: Randomly initialize the population
Step 3: Calculate fitness value for each particle
Step 4: Sort the particles in descending order of fitness value and save current best solutions in the repository Gbest.
Step 5: Update the population using horizontal crossover and vertical crossover.
Step 6: Repeat the steps until iteration number is reached.
Step 7: Select the best solution found in the above solving procedure.

Firefly Clustering:

A firefly yield flashing light as a signal to communicate with other fireflies, which also consequence on other fireflies flashing and flying movement. As an intelligent swarm method, firefly algorithm can be used for clustering analysis. Here firefly algorithm for keyword clustering. It indicated that firefly algorithm produced consistent and better performance in terms of time and optimality than other algorithms. After keywords are extracted, the user will give their choices. The keywords will be clustered according to their user choice. Finally, it will recommend to the user, so it improves user satisfaction query result [5].

Firefly Clustering Algorithm Steps:

Step 1: Set the initial value of randomization parameter, firefly attractiveness, media light absorption coefficient, population size and maximum generation number.
Step 2: Generate initial population randomly
Step 3: Evaluate the fitness function of all solution in the population
Step 4: update light velocity
Step 5: Rank the solutions and find the current best
Step 6: End

IV. EXPERIMENTAL RESULTS

In this work, the conversation transcript has been focused. The proposed method validated using data set from ELEA (Emergent Leader Analysis) Corpus. The text file is similar to meeting transcripts format.The ELEA Corpus consists of approximately ten hours of recorded and transcribed meetings, in English .This transcript is four party conversation (Speakers A-D) was submitted to the document recommendation system. By applying this proposed framework essential keywords are extracted from this Corpus. These inputs are tested by using both Existing and proposed methods. This will help for the accurate prediction of document keywords for recommendation.The performance of the proposed method is compared with the Existing methods. The results are shown that the proposed method is a better one. In the same way this proposed method will be implementing for other datasets also.

Metrics Considered for Evaluation:

The performance of the proposed framework is measured in terms of the quality measures namely Precision, Recall, F-Measure, Accuracy, RMSE and NDCG.

Precision

Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant. It is calculated as follows,

Precision = {Number of Relevant Keywords} ∩ {Number of Retrieved Keywords} / {Number of Retrieved Keywords}

Recall

Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. It is calculated as follows,

Recall = {Number of Relevant Keywords} ∩ {Number of Retrieved Keywords} / {Number of Relevant Keywords}

F Measure

F Measure computes both precision and recall as the test to compute the score. Here precision is the number of correct keywords divided by number of all returned keywords. Recall is the number of correct keywords divided by the number of keywords.

F = 2 * Precision * recall / Precision + recall

Accuracy

Accuracy calculates the proposition of correctly identified keywords, and estimated by using equation.

Accuracy = (TP + TN) / (TP+TN+FP+FN)

In respect of Keyword Extraction the terms are evaluated in below manner.

True Positive (TP) – Keyword correctly identified as a Keyword
True Negative (TN) – Non- Keyword correctly identified as non-keyword
False Positive (FP) – Non-Keyword incorrectly identified as a keyword
False Negative (FN) – Keyword incorrectly identified as non-Keyword

Root Mean Squared Error (RMSE)

It is the difference between keywords predicted by a system and the keywords actually observed from the input. It is estimated as,

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y)^2}{n}}.$$

Normalized Discounted Cumulative Gain (NDCG)

It measures the performance of a irrelevant keywords. IDCG is the maximum possible keywords and DCG is irrelevant keywords.

NDCG = DCG / IDCG

In Existing method using a diverse keyword extraction algorithm, diverse set of keywords extracts from the conversation. Then it shows some keyword list and text files and user select one of the text files that text files contain some clustered document related to the keywords. Clustering is done by using unsupervised document classification. Finally, it shows clustered document [2][12].

The results are shown in the Table 1. This table shows the performance comparison of the various techniques with proposed Swarm Intelligence technique .The experiments have been repeated for randomly shuffled conversation and the results are obtained using ELEA Corpus is shown in the Table 1.

	EXISTING METHOD				PROPOSED METHOD		
Technique used	**Precision**	**Recall**	**F-Measure**	**Technique used**	**Precision**	**Recall**	**F-Measure**
DKE(Diverse Keyword Extraction)	0.6383	0.5286	0.6796	SI (Swarm Intelligence)	0.8992	0.6832	0.7345
	0.707	0.6725	0.6893		0.9238	0.7709	0.7682
	0.6931	0.5498	0.5938		0.8345	0.5992	0.6285
	0.7522	0.6725	0.6131		0.8761	0.7015	0.6781
	0.7382	0.6983	0.7012		0.7930	0.7237	0.7930

Table 1: Precision, Recall and F-Measure

Table 1 tabulates the Precision and Recall achieved for various Techniques. In Table 2 shows the F-Measure value. It is observed from Fig 2 that the SI achieves the best F- Measure. It is also seen that the use of SI improves the performance of the proposed technique better than existing DKE.

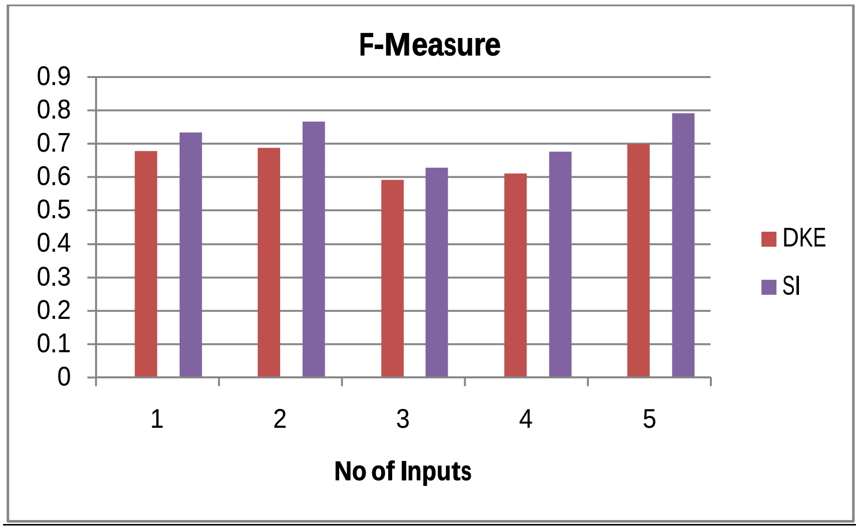

Fig 2: F-Measure Achieved

	EXISTING METHOD				PROPOSED METHOD		
Technique used	**Accuracy**	**RMSE**	**NDCG**	**Technique used**	**Accuracy**	**RMSE**	**NDCG**
DKE(Diverse Keyword Extraction)	67	0.0549	59	SI (Swarm Intelligence)	71	0.0321	62
	73	0.0472	78		76	0.0298	79
	75	0.0301	63		79	0.0231	65
	69	0.0416	67		72	0.274	79
	71	0.0317	71		74	0.0247	74

Table 2: Accuracy, RMSE, and NDCG obtained using SI Techniques

The result obtained from Accuracy, RMSE and NDCG obtained using SI techniques is shows in Table 2 and Fig 3 represents Accuracy and NGCG achieved. It is also seen that the use of SI improves the performance of the proposed technique better than existing DKE.

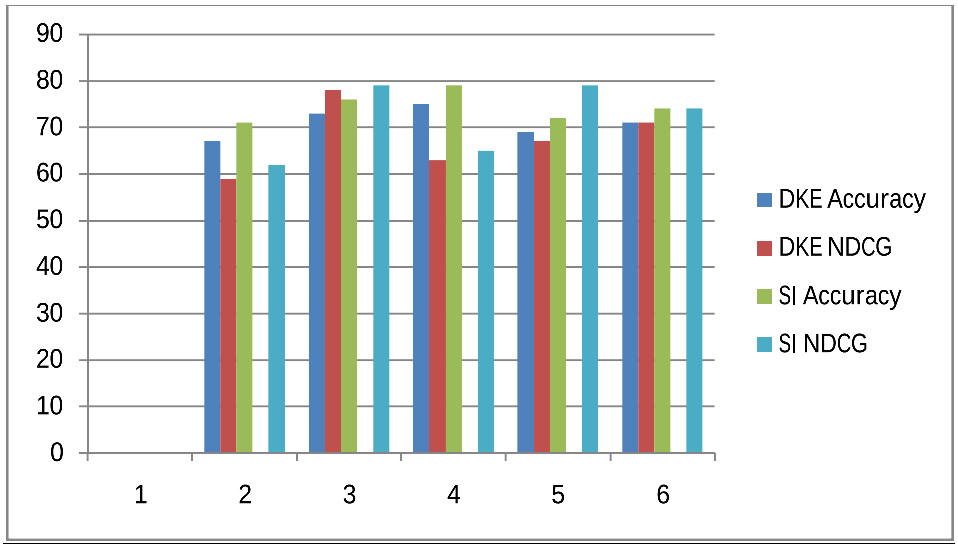

Fig 3: Accuracy and NDCG Achieved

V. CONCLUSION AND FUTURE WORK

The results showed that the proposed method performs better with the existing method. The proposed method provides searching and clustering keywords are easy to identify the user satisfaction based queries. This will produce good results in optimize space utilization, time consumption, cost efficiency, high accuracy compared to the Existing method. In the future, this will be tested with human users of the system within real-life meetings.

REFERENCES

[1] M. Habibie and A. Popescu-Belis, “Enforcing topic diversity in a document recommendation for conversations,” in Proc. 25th Int. Conf. Comput.Linguist. (Coling), 2014, pp. 588–599.
[2] M. Habibi and A. Popescu-Belis, “Diverse keyword extraction from conversations,” in Proc. 51st Annu.Meeting Assoc. Comput.Linguist., 2013, pp. 651–657.
[3] Jianxin Li, Chengfei Liu, “Context-Based Diversification for Keyword Queries over XML Data,” in IEEE Transactions on Knowledge and Data Engineering, vol. 27, 3 March 2015.
[4] T.Jo, “ Table Based KNN for Extracting Keywords”, in proc.IEEE Conference Publications in Computer Science and Information Engineering, 2016.
[5] Menaka S, Radha N,” Text Classification using Keyword Extraction Technique,” in Proc. Int. journal of Advanced research in computer science and software Engineering, vol 3, Issue 12, December 2013.
[6] Maryam Habibi and Andrei Popescu-Belis, “Keyword Extraction and Clustering for Document Recommendation in Conversations,” IEEE/ACM Transactions on audio, speech, and language processing, vol. 23, no. 4, April 2015.
[7] Timothy Man-Hung Siu Steve Lowe , Arthur Chan, “ Topic Modeling for Spoken Documents Using Only Phonetic Information,” IEEE conference on automatic speech recognition and understanding,pp.395-400, December 2011.
[8] M. Habibi and A. Popescu-Belis, “Using crowdsourcing to compare document recommendation strategies for conversations,” Workshop Recommendat. Utility Eval.: Beyond RMSE (RUE’11), pp. 15–20, 2012.
[9] Zhen Yue Shuguang Han Daqing He,” An Investigation of the Query Behavior in Task -based Collaborative Exploratory Web Search,” in Proc. 23th Int. Conf. Comput.Linguist. (Coling), 2013.
[10] A. Nenkova and K. McKeown, “A survey of text summarization techniques,” in Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. New York, NY, USA: Springer, 2012, ch. 3, pp. 43–76.
[11] P.Wu, X.Zhang, “A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion” in Proc.IEEE Conference publications in Computer Science Engineering, 2016.
[12]]D.Sowmya and J.I. Sheeba , “ Keyword Extraction using Particle Swarm Optimization” ,in Proc International Conference on Computational Modelling and Security - CMS 2016(Elsevier), ISSN: 1877 – 0509, 2016.