RESEARCH ARTICLE OPEN ACCESS

Construction of Keyword Extraction using Statistical Approaches and Document Clustering by Agglomerative method

R. Nagarajan*, Dr. P. Aruna**
*(Department of Computer Science & Engineering, Annamalai University, Annamalainagar)
** (Department of Computer Science & Engineering, Annamalai University, Annamalainagar)

ABSTRACT
Organize continuing growth of dynamic unstructured documents is the major challenge to the field experts. Handling of such unorganized documents causes more expensive. Clustering of such dynamic documents helps to reduce the cost. Document clustering by analysing the keywords of the documents is one the best method to organize the unstructured dynamic documents. Statistical analysis is the best adaptive method to extract the keywords from the documents. In this paper an algorithm was proposed to cluster the documents. It has two parts, first part extracts the keywords using statistical method and the second part construct the clusters by keyword using agglomerative method. This proposed algorithm gives more than 90% of accuracy.

Keywords– Agglomerative Method, Co-occurrences Statistical Information (CSI), Document Clustering, Similarity Measures, TF-ISF

I. Introduction

In this digital epoch, the tremendous increase of dynamic unstructured documents is unavoidable and should be organized in a good manner to use it cost effectively. This increase of unstructured documents raises the challenge to the field experts to use it effectively. Such documents are more informative and need to the fields those reveals around the data handling such as web search, machine learning ; Document Clustering is the most powerful method to solve the problem of organizing unstructured dynamic documents. There are various approaches available to cluster the documents. Clustering based on the concepts extraction is the straight and best method. Keywords help to extract the concept of the documents. Keywords are the words used in the documents, which summarises the concept of the documents. Extraction of the fruitful keywords from the bag of words is another challenge job. The basic assumptions that (i) authors of scientific articles choose their technical terms carefully; (ii) when different terms are used in the same articles it is therefore because the author is either recognizing or postulating some non-trivial relationship between their references; and (iii) if enough different authors appear to recognize the same relationship, then that relationship may be assumed to have some significance within the area of science concerned the keywords are extracted from the documents. In the first of part of the proposed algorithm extract the significant keywords by applying the statistical analysis on the bag of words, the second part of the algorithm deals the document clustering.

II. Related work

In this section we review previous work on document clustering algorithms and discuss how these algorithms measure up to the requirements of the Web domain. In [1] statistical feature extraction methods have been discussed and also framework for statistical keyword extraction is defined. In [2] survey of keyword extraction techniques have been presented and also deals the merits and demerits of the simple statistical approach, Linguistics Approach, Machine learning approach and other approaches like heuristic approach. In [3] a model for extracting keywords based on their relatedness weight among the entire text terms has been discussed and strength of the terms are evaluated by semantic similarity. In [4] different ways to structure a textual document for keyword extraction, different domain independent keyword extraction methods, and the impact of the number of keywords on the incremental clustering quality are analyzed and a framework for domain independent statistical keyword extraction is introduced. In [5] hybrid keyword extraction method based on TF and semantic strategies have been discussed and also semantic strategies were introduced to filter the dependent words and remove the synonyms. In [6] suffix tree clustering algorithm has been discussed and the authors also create an application that use this algorithm in the process of clustering, and search of clustered documents. In[7] novel *down-top* incremental conceptual hierarchical text clustering approach using CFu-tree (ICHTC-CF) representation has been discussed. In [8] variety of distance functions and similarity measures are compared and analyzed, the effectiveness of these measures in partition clustering for text document

datasets has been discussed and also the results are compared with standard K-means algorithm. In [9] different agglomerative algorithms based on the evaluation of the clusters quality produced by different hierarchical agglomerative clustering algorithms using different criterion functions for the problem of clustering medical documents has been discussed. In[10] text document space dimension reduction in text document retrieval by agglomerative clustering and Hebbian-type neural network has been discussed.

III. Methodology

IV. 3.1 Feature Extraction

The first part of our proposed algorithm deals with the keyword extraction using statistical analysis on the words of the documents. The steps involved in the proposed keyword extraction algorithm

- ¬ Pre-processing of the documents.
- ¬ Construction of sentences Vs keyword matrix.
- ¬ Calculate the weight of the words of the documents.
- ¬ Rank the words based on their weight.
- ¬ Find out the keywords based on the higher weights.

Step 1:
For each document D
do
Begin
Step 2: removal of stop words, stemming words, and removal of unnecessary characters and word simplification.

Step 3 : Construction of Sentences Vs Words Matrix
i. extract sentences from the documents and labeled as DS_i
ii. extract words from each sentence DS_i and stored in a Sentence DS_iW_j array.
iii. construct the Sentences Vs words matrix using Sentences DS_iW_j array.

Step 4 : calculate the words weight using the following statistical approaches
i. Most frequency words
ii. Term Frequency - Inverse Sentence Frequency
iii. CSI Measure (Co-occurrence statistical information) for noise removal from co-words construction

Step 5 : Extraction of keyword from higher weight words
End

In the preprocessing stage, the stop words and the unnecessary words are removed, then the stemming of the words are done and finally the words are simplified. All the sentences are extracted from the preprocessed document and labeled as DSi, words in the sentences are extracted with their frequency and their sentence label. To find out the keyword of the documents, sentences Vs words matrix is constructed. Table-1 shows the sentence-word matrix.

Table -1
Sentences-Words matrix

Sentences/ Words	W_1	W_2	W_3	...	W_j
S_1	S_1W_1	S_1W_2	S_1W_3	...	S_1W_j
S_2	S_2W_1	S_2W_2	S_2W_3	...	S_2W_j
S_3	S_3W_1	S_3W_2	S_3W_3	...	S_3W_j
...	...	...	...	...	...
S_i	S_iW_1	S_iW_2	$S_i W_3$	...	S_iW_j

In Table-1, each row corresponds to a sentence of a documents and column represents word. The set of sentences are represented as S = {S1,S2,S3,...Si} and set of words are represented as W = {W1,W2,W3....Wj}. The value 1 is assigned to a cell (S1W1) if the word occurs in that sentence and the value 0 is assigned otherwise. To compute the word weight three statistical methods are used. i) Higher Frequency words ii) Term frequency – Inverse Sentence Frequency iii) Co-occurrence Statistical Information.

i) Higher Frequent words (HF):

Higher frequent is the basic statistical measure, it just extract keywords straightly from higher frequency words. The word weight is calculated by counting the number of occurrence of the word in the Sentence-word matrix. i.e

$$HFW(W_j) = \sum_{S_{i \in S}} W_j S_i$$

where W_j is j^{th} word in a document , and S_i is the i^{th} Sentence in a document.

ii) Term Frequency – Inverse Sentence Frequency(TF-ISF) :

The TF-ISF is the another statistical measure to find out the weightage of the words in the documents. It finds out the weight of the word according to its frequency and its distribution through the sentences of the document. The weight of the word is given by

$$TF\text{-}ISF(W_j) = Frequency(W_j) * \log\left(\frac{|S|}{Frequency\ (W_j)}\right)$$

Where W_j is the j^{th} word in the document, In this method the weight of the word is less when it occurs more number of sentences. That is it should be

identified as a common word and it will not helps to identifies the concept of the document.

iii) Co-occurrences Statistical Information(CSI):

CSI is another statistical measure to find out the weight of the words in the documents using $\chi 2$ measure. It also find out the word which has more frequency but not such important word to find the concept of the document. $\chi 2$ measure calculate the deviation of the observed frequencies from the expected frequencies. The $\chi 2$ measure of word (wj) is given by

$$CSIW(w_j)=\chi^2(W_j)= \sum_{W_j \in S} \frac{(CO-OCCUR(W_j,W_K)-CO-OCCUR(W_j)p(W_k))^2}{co-occur(W_j)p(W_k)}$$

in which p(wk) is the probability of the word wk occurs in the sentence-term matrix and co-occur(wj) is the total number of co-occurrences of the term wj with terms wk € W.

In this case, co-occur(wj,wk) corresponds to the observed frequency and co_occur(wj)p(wk) corresponds to the expected frequency. Generally, all documents are composed by sentences with variable length, words used in a lengthy sentences tends to co-occur with more words used in that sentence. So, our keyword extraction approaches identify the keywords erroneously, to rectify such false identification, the CSI measure redefined p(wj) as the sum of the total number of words in sentences where wk appears divided by the total number of words in the document, co-occur(wj) as the total number of words in sentences where wj appears. Moreover, the value of $\chi 2$ measure can be influenced by non important but adjunct terms. To make the method more robust to this type of situation, the authors of the CSI measure subtracts from the $\chi 2$ (wj) the maximum $\chi 2$ value for any wk € W, i.e.:

$$CSIW(W_j) = \chi^2(w_j) - \operatorname{argmax}_{W_k \in W} \left\{ \frac{(freq(W_j W_k) - nW_j pW_k)^2}{nW_j pW_k} \right\}$$

By comparing the weights of the words derived from the three statistical approaches, the common accepted top weighted keywords are identified with their documents and labeled as a keywords of the document.

V. Document Clustering

5.1 Clustering process

Clustering is the process of grouping similar documents into sub sets based on their aspects and each subset is a cluster, i.e in a cluster the document are similar to each other. Unlike classification, clustering do not need any training data. Because of this nature, it is better suited to cluster unsupervised documents. In this proposed method, agglomerative hierarchical clustering approach is followed to construct clusters. It is down-top method. Our proposed methods starts by leasing each document be a cluster and iteratively either merges clusters into larger clusters or splits the cluster. Merging process is followed when two clusters are closed to each other's according to the similarity measures, inversely splitting process is followed when two clusters are far away, and this iteration process is continued till the termination constraint reached.

5.2 Similarity Measure

In this section, the distance between the documents are calculated with the features of the documents derived,

$$Dist(D_iF_i, D_jF_j) = \sqrt{\sum_{K=1}^{m} |d_{ik} - d_{jk}|^2}$$

where $i \neq j$

DiFi and DjFj represents the features of the documents Di and Dj respectively and taken as two individual clusters Ci and Cj, If the distance between the two clusters is maximum value, that show there is no common features between them. Inversely the distance between two clusters is minimum value when two clusters have common features.

5.3 Merging of clusters

The range of values allowed for the distance is 0 to √2. To normalize the similarity measuring values The similarity between the clusters i and j is defined as

$$Sim(F_i,F_j) = \sqrt{\frac{\sum_{K=1}^{m} |d_{ik} - d_{jk}|^2}{2}} \text{ where } i \neq j$$

By the similarity measure, the value is 0 assigned when the similarity measure between the two documents is far away and 1 when the distance between the two documents is closer. Closet pair of clusters are merged together and form one larger cluster.

$$Cluster(C_i, C_j) \leftrightarrow \max\{Sim(D_iF_i, D_jF_j) \quad i \neq j \text{ and } Sim(D_iF_i, D_jF_j \geq \Theta$$

where Ci and Cj are clusters can be merged, Sim(DiFi,DjFj) is the similarity between Ci and Cj, Fi and Fj are the features of Ci and Cj, respectively.

Step 1 : Initialize T (Tree)

Step 2 : for each D_i in a document set

Step i : $c_i \leftarrow$ preprocessed(D_i)
Step ii : add c_i to T as a separate node
Step iii : Repeat for each pair of clusters C_j and C_k in T
if C_j and C_k are the closest pair of clusters in T
then
merge(C_j, C_k)
compute cluster feature vectors changed
Else
split (C_j, C_k)
compute cluster feature vectors changed
Endif
Until all clusters are not changed
End for

Step 3 : Return T

VI. Experimental Analysis

To validate our proposed algorithm, we conduct the experiments on the sample data. 40 documents are considered for experiment. Initially our keyword extraction portion of algorithm is applied. In the first stage 3800 words are extracted from sample of documents, after preprocessing the documents, 1428 unique words and 19420 sentences are extracted. Then the algorithm constructs the Sentences Vs Words matrix for 19420 sentences as rows and 1228 words as column to test our TF-ISF statistical approach. The cell value is filled with 1 if the word occurs in the sentence and 0 otherwise.

To weight the words extracted, we apply three statistical approaches, First, Most Frequent words-it just identifies the higher order frequency words as the keywords irrespective of other measuring methods, for the 1428 unique words, it identifies 828 words as the keywords (threshold value 5). Term Frequency-Inverse Sentence Frequency (TF-ISF), second statistical approach, it find out the weight of the word according to not only its frequency but also its distribution through the sentences of the document. By this TF-ISF, the words with high frequency value but truly not much important to help to identify the concept of the document are eliminated, because those words may occur in more number of sentences. Finally 710 words are identified by this TF-ISF statistical approach.

Co-occurrences Statistical Information(CSI)-third statistical measure to find out the weight of the words in the documents using χ^2 measure. It also finds out the word which has more frequency and less presence in the sentences but not much important word to find the concept of the document. χ^2 measure the deviation of the observed frequencies from the expected frequencies. By applying the Co-occurrences Statistical Information (CSI) approach, 640 words are extracted as keyword of the sample data. By Comparing the words resulted from the three statistical approaches 590 words were labeled as keywords of our sample documents. The following Table-2 shows the results obtained from the three statistical approaches.

Table-2
Keywords extracted by applying statistical approaches

R	MF	TF-ISF	CSI
1	Data Mining	Data Mining	Data Mining
2	Machine Learning	Machine Learning	Machine Learning
3	Image Mining	Image Mining	Image Mining
4	Recognition	Database	Data encryption
5	Segmentation	Data encryption	Database
6	Database	Graphics	Data set
7	data encryption	Pre-processing	Pre-processing
8	data compression	Security System	Data compression
9	Data set	Information Retrieval	Segmentation
10	Pre-processing	Segmentation	Data encryption
11	Graphics	Data set	Graphics
12	Information Retrieval	data compression	Security System
12	Security system	Router	Information Retrieval
13	Hub	Neural Networks	SVM
14	Router	SVM	Hub
15	Neural networks	Hub	Router
16	SVM	Clustering	Clustering
17	Clustering	Recognition	Recognition

From the Table-2 values, the top ranked keyword of 14 documents are derived and displayed in the following Table-3

Table-3
Documents keyword representation

Doc.Id	Features/Keywords Extracted
D1	{ data mining, dataset, preprocessing ,information retrieval, machine learning}
D2	{ image mining, graphics, recognition, segmentation}
D3	database, data encryption, data compression, data mining}
D4	{data mining, preprocessing, dataset, machine learning}
D5	{database, data compression, data encryption}

D6	{image mining, recognition, segmentation}
D7	security system, hub, router}
D8	{database, data encryption, data compression}
D9	{ image mining, recognition, graphics, segmentation}
D10	{neural network, SVM, clustering}
D11	{data mining, machine learning, dataset}
D12	{data mining, preprocessing, machine learning, information retrieval}
D13	{image mining, recognition, segmentation}
D14	{database, data encryption, data compression}

After extracting the features, the distance between the documents are measured by the similarity measure, the following matrix shows the distance between sample documents as numerical values. The values ranged from 0 to 1.

To create clusters, our second part of the algorithm executes by assuming the document D1 as our first node, the value of D1 and D2, D1 and D3, D1 and D4, D1 and D5… D1 and D14 are compared, the values closer to 1 indicates that those documents deals same concepts, the distance values between D1 to D4,D11,D12 are 0.9, 0.8, 0.7 respectively, which are closer to value 1 and forms the cluster C1{D1,D4,D11,D12}. Likewise the values between D2 to D6,D9,D13 are 0.8, 0.9 and 0.9 respectively shows that the second cluster C2 merges the documents D6,D9 and D13 with D2 forms C2{D2,D6,D9,D13}. The values between D3 to D5,D8 and D14 are 0.9, 0.9 and 0.8 forms the third cluster C3{D3,D5,D8,D14}. It is also identified from the table value of D7 and D10 with all the other documents are far away that is so smaller than 1, which shows that documents the concepts of D7 and D10 are far away the concepts of other documents taken for testing. Finally three number clusters are formed from our testing data set.

Figure-1 shows the distance values between the documents C1\D1, D4, D11, D12 C2\D2, D6, D9, D13 C3\ D3, D5, D8, D14

VII. Conclusion

In this paper, we proposed an algorithm for feature extraction and document clustering, three statistical approaches are applied to the words of the documents. Because of the different base natures of the statistical approaches the fake features are eliminated, the values arrived by the statistical approaches are compared and the top ranked words are labeled as the keywords or features of the documents. Then, the distance between the documents are calculated using the features extracted. By applying the similarity measuring the close concept documents forms the clusters. Summarily, according to the experimental results, the document clustering method proposed in this paper handles the unstructured unlabelled documents effectively.

References

[1] Rafael Geraldeli Rossi, Ricardo Marcondes Marcacini, Solange Oliveira Rezende, "Analysis of Statistical Keyword Extraction Methods for Incremental Clustering", 2013.

[2] Jasmeen Kaur and Vishal Gupta," Effective Approaches For Extraction Of Keywords", *IJCSI International Journal of Computer Science Issues", Vol. 7, Issue 6, 2010.*

[3] Mohamed H. Haggag, "Keyword Extraction using Semantic Analysis", *International Journal of Computer Applications, Volume 61– No.1, 2013.*

[4] Rafael Geraldeli Rossi , Ricardo Marcondes Marcacini , Solange Oliveira Rezende," Analysis Of Domain Independent statistical keyword extraction methods for Incremental Clustering", *Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 12, Iss. 1, pp.17 37, 2014.*

[5] S. Wang, M. Y. Wang, J. Zheng, K. Zheng, "A Hybrid Keyword Extraction Method Based on TF and Semantic Strategies for Chinese Document", *Applied Mechanics and Materials, Vols.635-637, pp.1476-1479, 2014.*

[6] Milos Ilic, Petar Spalevic, Mladen Veinovic," Suffix Tree Clustering - Data mining algorithm", *ERK2014, Portoroz, B:15-18, 2014.*

[7] Tao Peng and Lu Liu, "A novel incremental conceptual hierarchical text clustering methodUsing CFu-tree", *Applied Soft Computing, Vol. 27, pp. 269-278, 2015.*

[8] Anna Huang, "Similarity Measures for Text Document Clustering", *New Zealand Computer Science Research Student Conference", pp.49-56, 2008.*

[9] Fathi H. Saad ,Omer I. E. Mohamed , and Rafa E. Al-Qutaish, "Comparison of Hierarchical Agglomerative Algorithms for Clustering Medical Documents", *International Journal of Software Engineering & Applications, Vol.3, No.3, 2012.*

[10] Gopal Patidar, Anju Singh and Divakar Singh, "An Approach for Document Clustering using Agglomerative Clustering and Hebbian-type Neural Network*", International Journal of Computer Applications, Vol.75, No.9, 2013.*

Figure -1
Document Vs Document Distance Matrix

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14
D1	X	0.3	0.2	0.9	0	0.3	0.2	0.4	0.4	0	0.8	0.7	0.2	0
D2	0.3	x	0.3	0.2	0	0.8	0	0.1	0.9	0	0.2	0.3	0.9	0
D3	0.2	0.3	x	0.3	0.9	0.2	0.1	0.9	0.1	0	0.4	0.4	0.4	0.8
D4	0.9	0.2	0.3	x	0.3	0.1	0	0.3	0.1	0	0.7	0.8	0.3	0.2
D5	0	0	0.9	0.3	x	0.2	0	0.2	0.2	0	0.4	0.3	0.4	0.7
D6	0.3	0.8	0.2	0.1	0.2	X	0	0.2	0.7	0	0.3	0.3	0.8	0.4
D7	0.2	0.1	0	0	0	0	x	0.1	0	0	0	0	0	0
D8	0.4	0.1	0.9	0.3	0.2	0.2	0.1	x	0.4	0	0.3	0.4	0.3	0.7
D9	0.4	0.9	0.1	0.1	0.2	0.7	0	0.4	x	0	0.4	0.3	0.8	0.4
D10	0	0	0	0	0	0	0	0	0	x	0	0	0	0
D11	0.8	0.2	0.4	0.7	0.4	0.3	0	0.3	0.4	0	x	0.7	0.3	0.4
D12	0.7	0.3	0.4	0.8	0.3	0.3	0	0.4	0.3	0	0.7	x	0.3	0.3
D13	0.2	0.9	0.4	0.3	0.4	0.8	0	0.3	0.8	0	0.3	0.3	x	0.2
D14	0	0	0.8	0.2	0.7	0.4	0	0.7	0.4	0	0.4	0.3	0.2	x