International Journal of Computer Applications Technology and Research
Volume 3– Issue 3, 165 - 168, 2014, ISSN: 2319–8656

Re-enactment of Newspaper Articles

Thilagavathi .N
Sri ManakulaVinayagar
Engineering College
Pudhucherry, India

Archanaa S.R
Sri ManakulaVinayagar
Engineering College
Pudhucherry, India

Lavanya.K
Sri ManakulaVinayagar
Engineering College
Pudhucherry, India

Valarmathi.S
Sri ManakulaVinayagar
Engineering College
Pudhucherry, India

Abstract: Every document that we use has become digitized which makes a great way to save, retrieve and protect documents. They are digitized to have a backup for most paper work .Digitization is found to be more important since everything is going paper free. Digitization of newspaper contributes greatly to preservation and access to newspaper archives. Our paper provides an integrated mechanism that involves document image analysis and k means clustering algorithm to digitize news articles and provide an efficient retrieval of user requested news article. In first stage the news article is segmented from newspaper and pre-processed. In the second stage the pre-processed news articles are clustered by K-means clustering algorithm and key words are extracted for each cluster. The third stage involves selection of cluster containing key phrase given by user and providing the user with requested news article.

Keywords: Page segmentation, TF-IDF weighting, Cosine similarity, Clustering, K-Means algorithm, Keyword Extraction.

1. INTRODUCTION

Document digitization plays a vital role in electronic publishing of newspaper. Digitization of newspaper has become very essential to protect historical news articles, easy storage and efficient retrieval of news articles when needed. In order to obtain the above functionalities, the digitized newspaper need to be powered up with algorithms for document image analysis , efficient storage and retrieval to avoid the ambiguousness during the retrieval of specific news article. Moreover transferring the news article into the system by hand consumes more time and human resource. Thus there is a need for an automated system to obtain the above functionalities.

The basic unit of newspaper is composed of news articles. Document Image Analysis is done to obtain the articles from each and every section of the newspaper one by one. This task is very challenging since it needs to consider the syntactic and semantic information of the blocks of content in every news article. Using the syntactic and semantic information from the image analysis, the newspaper is segmented into individual news article. The content of the each segmented news article is converted into a word file and stored into the database using clustering algorithm. Clustering of news articles involves grouping of news articles into clusters, where they share common properties and keywords. Here, we implement k-means clustering algorithm which is suitable for huge data set like digitized newspapers of years and years. Further, the keyword for each cluster is determined for the efficient retrieval of the required article based on search phrase provided by the user.

2. RELATED WORKS

There are many researches done on newspaper digitization, storage of digitized newspaper and efficient retrieval of them. Most of the existing system does not combine best approach for all three processes together. LiangcaiGao et al. proposed a method to reconstruct Chinese newspaper [5] by accomplishing several tasks such a article body grouping, reading order detection, title-body association and linking scattered article by travelling salesman problem (TSP) and Max-Min Ant System (MMAS). In order to increase the efficiency of MMAS a level based pheromone mechanism is done. It includes two subtask enactment of news article in reverse order by detecting reading order and then using the content continuity to aggregate the text blocks. This method is time prone since it involves semantic analysis of the newspaper content to separate the news article from newspaper. Fu Chang et al. established an approach for layout analysis using adaptive regrouping strategy for Chinese document [2]. This method is specific for Chinese documents that involve horizontal as well as vertical text lines. Wei-Yuan Chen uses an adaptive segmentation method to extract text blocks from colored journals [1] which involves RLSA (run-length smoothing algorithm).This approach needs improvement to adjust the segmentation of non-uniformly colored character from background with complex color.

Osama Abu Abbas proposed a comparison between the four major clustering algorithm k-means algorithm, hierarchical clustering algorithm, self-organization map (SOM) algorithm and expectation maximization algorithm [3]. These algorithms were selected for comparison based on their popularity, flexibility, applicability and handling high dimensionality. These algorithms was compared based on size of dataset, number of clusters, type of dataset the algorithms are going to handle and type of software those algorithm is to be implemented. The result shows that the k-means clustering algorithm is known to be efficient in clustering large data sets. The k-means algorithm allows discovery of clusters from subspaces by identifying the weights of its variables and it is also efficient in identifying noise variables in data. K-means algorithm is suitable for variable selection in data mining. FarzadFarahmandnia proposes a method for automatic key word extraction in text mining using WordNet[4]. By this method the text files are normalized by TDIDF algorithm and preprocessed to remove stop words. Then each word in the text file are hierarchically structured in WordNet dictionary .In order to avoid ambiguities between search words in hypernym search, comparison of every pair of words in document is done. This is done by determining the distance between the two words which is calculated by number of edges between node nodes with search word. Thus words with much closer distance will be chosen as key words for the text document. This paper proposes an approach to segmentation of news article from newspaper, clustering of news article based on its content and assigning labels for each cluster using WordNet.

3. SYSTEM ARCHITECTURE

The proposed system uses scanned newspaper images as input. A newspaper page image contains many articles. These articles are segmented from newspaper using the method, article segmentation by which each news article is made as a text file. These text file is preprocessed to remove stop words and stem words. In order to compare the text documents to compute similarity we perform TF-IDF weighting and cosine similarity. Based on the similarity between the documents, K-means algorithm is used. It is done to cluster documents that express maximum similarity. Keywords for each cluster are extracted to enhance searching. When the user query for a news article the requested news article is retrieved based on key word matching, post processing method involving WordNet.

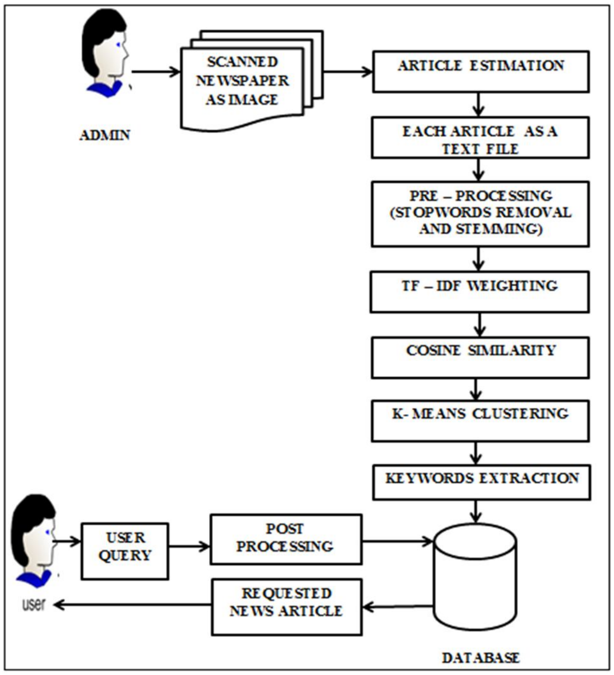

Figure 1. Architecture diagram of Re-Enactment of newspaper article

4. RESEARCH PROPOSAL

4.1 Page Segmentation

To start page segmentation to obtain news article from scanned newspaper image the first essential element to be identified are horizontal and vertical foreground lines. They indicated the boundary of the news article in a newspaper. In order to identify the boundary, binary image of newspaper is transferred into grayscale image. The grayscale image is sub-sampled with respect to foreground pixel. From the result, we obtain two images by assigning all foreground pixel with the length of vertical or length of horizontal line. Thus the horizontal and vertical line needs to be identified are resulted. It is identified since the sub-sampled gray scale image is applied with a condition that is to obtain only pixels whose length or width is larger or smaller respectively than the threshold. Thus the pixel featuring only the horizontal and vertical boundaries of the article is obtained as result. The final stage of segmentation is to extract text from the segmented image. The result of sub-sampling with respect to background pixels are used in order to avoid extracting text from neighbor block. Each block of image is given as an input to OCR (Optical Character Recognizer) which converts each article into a text document.

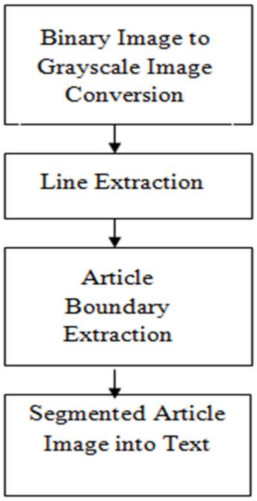
Binary Image to Grayscale Image Conversion
Line Extraction
Article Boundary Extraction
Segmented Article Image into Text

Figure 2. Segmentation of news article from newspaper

4.2 Pre-Processing Of Documents

For every article in the scanned newspaper, a word document is created. Pre-processing of these word file has to be done to prune words from the document with poor information. It optimizes the keyword list that contain list of terms in the document. It involves removal of stop words and stemming words. Pronouns, preposition, conjunction and punctuations carry no meaning as keywords are to be removed in pre-processing. The words in the document are listed out and if it is present in the list of stop words that has been pre-defined in our method, they are removed. This is followed by removal of stemming words. It involves finding variant for a word and replaces it with main word. This is done with the help of WordNet.

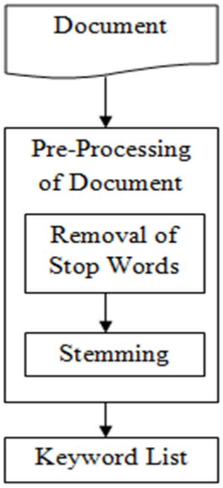

Figure 3. Steps involved in pre-processing

4.1.1 Tf-Idf Weighting

Before applying clustering algorithm on a set of news articles as word documents, for comparing the documents, they must be converted into vector representation. The pre-processed document must be represented with TF-IDF score. TF-IDF stands for Term Frequency-Inverse Document Frequency which results the importance of a term among the document. Term frequency is calculated by dividing the number of occurrences of a word in its document by total number of word in the document. It is a normalized frequency. Inverse document frequency is calculated by taking log of number of documents to be clustered divided by number of documents containing the term. It gives higher weight to rare items. Multiplying the two metrics together give TF-TDF weighting which gives importance to terms frequent in the particular document weighted for clustering and rare among the documents that are clustered.

Tf –Idf (term, document)
=Tf (term, document)*Idf (term)

Where Tf is term frequency
Idf is inverse document frequency

4.1.2 Cosine Similarity

As a result of TD-IDF weighting, we have represented each news article in the form of word document as vector models. Next step is to find the similarity between the documents. In our method cosine similarity is used to obtain the distance (similarity) between two documents. It is computed by dividing the dot product of two vectors by the product of their magnitudes. This defines equidimensionality and element- wise comparability of document vectors in vector space. The cosine angle is a good indicator of similarity between the two vectors of the documents.

$$\text{Cosine similarity (vec\_A, vec\_B)} = \frac{\text{Dot Product (vec\_A, vec\_B)}}{|\text{vec A}| * |\text{vec B}|}$$

Where vec_A is vector model of document A
vec_B is vector model of document B

4.3 Clustering

The news article documents are to be clustered to improve the results of information retrieval system in terms of precision or recall. This provides better filtered and adequate result to the user. Clustering methods are made into generic categories: hierarchical agglomerative and partitional clustering. Hierarchical clustering is of two types. One forms a sequence of partition in data that leads n clusters from single cluster (divisive) and another merge clusters based on similarity between clusters (agglomerative). The divisive algorithm starts up with each data point as a cluster. Then it merges the tree node that shares certain degree of similarity. Thus it needs either cluster similarity or distance measure to split or merge data of different cluster. Agglomerative algorithm involves pair wise joining of clusters. Hierarchical clustering algorithms face difficulties in handling different sized cluster and not suitable for large sized data. Thus we prefer partition algorithm which suits large set of data.

Partitional algorithm defines the number of clusters initially, let k and evaluate the data at once such that sum of distance over their cluster center is minimal. Unlike hierarchical clustering, partitional clustering involves single level division of data. There are various types of partitional clustering algorithms: k-means, k-median and k-medoids. These algorithms differ by the approach of defining cluster centers and not how they represent the clusters k-means algorithm defines its center as mean data vector averaged over all data nodes in the cluster. In k-median the median is calculated for each dimension in data vector. In k-medoids the cluster center is defined as an item with smallest sum of distances to other items in the cluster.

4.1.3 K-Means Clustering

K-means algorithm is an unsupervised learning algorithm which is much efficient than other partition algorithm with better initial centroids .It aims to partition n documents into k clusters in which each document belongs to cluster with nearest mean that is, it groups similar document where each group is known as a cluster. Document in each group establish maximum similarity within its group and maximum diversity with other groups.

Step 1: Initialize parameter k, number of cluster centroids based on number of cluster needed.

Step 2: Data points are assigned to the closest cluster based on the cosine similarity.

Step 3: The position of the centroids are recomputed after assigning all data points are assigned to the cluster.

Step 4: Step 2 and 3 are repeated until cluster converge.

Initially the user has to specify the value of k, desired number of cluster centers. Each data point is assigned to the nearest centroid. Set of points assigned to each centroid is known as cluster. When data points are added the centroid for the cluster is updated based on the added data points

4.4 Keyword Extraction And News Article Retrieval

After the clusters are formed by the clustering algorithm, keywords for each cluster have to be defined for each cluster. In order to define key words list for each cluster, we first select the frequent terms in the cluster by setting threshold. The resultant list is fed to the WordNet, an electronic lexical database that describe each English word as noun, adverb, adjective and verb. It also describe the semantic relationship between the word that is, it is whether its synonym or hyponym. WordNet collect the noun candidates from the keyword list of the cluster and consolidate the set of synonym and hypernym words. Thus keywords and the related synonyms and hyponyms are defined for each cluster. Thus the user queries the cluster database with user defined key phrase. The words in the key phrase are compared with the keyword list of cluster. The cluster with which the key phrase matches is said to contain the required news article.

5 CONCLUSION

Re-enactment of newspaper article proposes an approach to segment news article from newspaper and convert those article into word files. These word files are pre-processed to remove stop words and stemming. This pre-processed word file is converted into vector form by means of TF-IDF weighting. Each document is represented by means of a vector. The similarity between the documents is found out by means of cosine similarity. The documents with more similarity are clustered by means of K-means algorithm. For each cluster formed by k-means algorithm keyword list are generated for making retrieval of article based on user queries efficient.

6 REFERENCES

[1] Wei-Yuan et al, Adaptive Page Segmentation for Color Technical Journals' Cover Image, Image and Vision Computing, 16(1998) 855-877, Elsevier Publication.

[2] Fu Chang et al, Chinese Document Layout Analysis Using an Adaptive Regrouping Strategy, Pattern Recognition 38(2005) 261-271,Pergamon Publication.

[3] Osama Abu Abbas et al, Comparisons between Data Clustering Algorithms, volume 5, No.3, July 2008, The International Arab Journal of Information Technology.

[4] FarzadFarahmandnia et al, A Novel Approach for Keyword Extraction in Learning Object Using Text Mining and WordNet, Volume 03, Issue 1(2013) 01-06,Global Journal of Information Technology.

[5] LiangcaiGao et.al, Newspaper Article Reconstruction Using Ant Colony Optimization and Bipartite Graph, Applied Soft Computing 13(2013) 3033-3046, Elsevier publication.