International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

Information Retrieval & Text Analytics Using Artificial Intelligence

Lavanya N[1], Dr Mahesh K Kaluti[2]

[1]M Tech, CSE, Dept. of Computer Science and Engineering, PESCE, Mandya
[2]Assistant professor, CSE, Dept. of Computer Science and Engineering, PESCE, Mandya

---

Abstract—*Information Retrieval is a key technology for knowledge management. The information seeker formulates a query trying to describe the information in need. There is a need for technologies that will allow efficient and effective processing of huge datasets. An effective document processing system must be able to recognize structured and semi structured forms that is written by different people's handwriting. In this paper, we use an OCR method and K means clustering algorithm to read handwritten documents by computer system and provide required information using efficient Information Retrieval approaches using Artificial Intelligence.*

Keywords— Text Analytics, Information Retrieval ,OCR, Artificial Neural networks, Information Extraction.

Introduction:

This Artificial Intelligence (AI) is defined as intelligence exhibited by an artificial entity to solve complex problems and such a system is generally assumed to be a computer or machine. Artificial Intelligence is an integration of computer science and physiology Intelligence in simple language is the computational part of the ability to achieve goals in the world. Intelligence is the ability to think to imagine creating memorizing and understanding, recognizing patterns, making choices adapting to change and learn from experience. Artificial intelligence concerned with making computers behave like humans more human like fashion and in much less time then a human takes. Hence it is called as Artificial Intelligence.

Text Analytics is the most recent name given to Natural Language Understanding, Data and Text Mining. In the last few years a new name has gained popularity, Big Data, to refer mainly to unstructured text (or other information sources), more often in the commercial rather than the academic area, probably because unstructured free text accounts for 80% in a business context, including tweets, blogs, wikis and surveys .Text Analytics is the discovery of new, previously unknown information, by automatically extracting information from different written resources.

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for metadata that describe data, and for databases of texts, images or sounds.

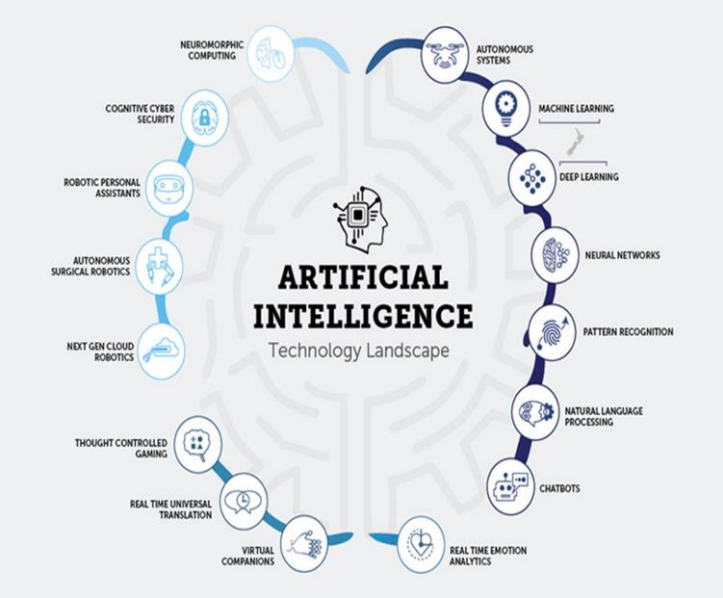

Fig: Overview Of Artificial Intelligence

Optical Character Recognition (OCR) is a process by which we convert printed document or scanned page to ASCII character that a computer can recognize. The document image itself can be either machine printed or handwritten, or the combination of two. Computer system equipped with such an OCR system can improve the speed of input operation and de-crease some possible human errors. Difference in font and sizes makes recognition task difficult if preprocessing, feature extraction and recognition are not robust and width of the stroke is also a factor that affects recognition. Therefore, a good character recognition approach must eliminate the noise after reading binary image data, smooth the image for better recognition, extract features efficiently and classify patterns.

Artificial Neural Network approach has been used for classification and recognition. It is a computational model widely used in situation where the problem is complex and data is subject to statistical variation. Training and recognition phase of the ANN has been per-formed using conventional back propagation algorithm with two hidden layers. The architecture of a neural network determines how a neural network transfers its input into output. This transfer can be viewed as a computation.

Methodology:

OCR System:

Following steps have been followed in the OCR system

A. Preprocessing
B. Segmentation
C. Feature Extraction

Preprocessing

In the OCR system, text digitization is done by a flatbed scanner. The digitized images are usually in gray tone, and for a clear document, a simple histogram based threshold approach is sufficient for converting them to two tone images.

The histogram of gray values of the pixels shows two prominent peaks, and a middle gray value located between the peaks is a good choice for threshold.

i. **Thresholding:** This is the first preprocessing stage. Gray or color image is taken as a input to convert into binary image, based on threshold. The output image obtained replaces all pixels in the input image with luminance greater than threshold level with the value 1 (white) and replaces all other pixels with the value 0 (black). Inverted binary image performs inverse operation, replaces 1‟s with 0‟s and 0‟s with 1‟s.In proposed algorithm 0.9 threshold is found good for conversion.

ii. **Noise Reduction And Normalization:** Binary image exists some noise which has to be removed for further operation that is segmentation and features extraction. Different Morphological operations are required for noise reduction. Morphologically open binary image removes small objects or all connected components fewer than threshold pixels, brings accuracy in features extraction. Clean operation removes isolated pixels individual 1s that are surrounded by 0s that is small dots from image which is beneficial to bring accuracy in segmentation. After noise reduction unwanted 0‟s around the character are removed by cropping and image is normalized and used for features extraction after segmentation.

Segmentation

The Segmentation is one of the most important phases of OCR system. Segmentation subdivides an image into its constituent regions or objects.

i. **Line segmentation:** The text lines are detected by finding the valleys of the projection profile computed by a row-wise sum of black pixels. This profile shows large values for the headline of the individual text line. The position between two consecutive headlines, where the projection profile height is minimum, denotes the boundary between two text lines.

ii. **Word segmentation:** To segmenting the text line image into words, compute vertical projection profiles. The projection profile is the histogram of the image. In the profile, the zero valley peaks represent the word space. Segmented line from first stage is taken as input for second stage that is word segmentation. Line is then segmented into word in this stage.

iii. **Character Segmentation:** Character segmentation is done after the individual words are identified. To extract character from word removal of headline is essential. For this first horizontal projection of individual word is computed and the rows having highest projection is consider as a headline and removed for further character segmentation. After removal vertical projections of individual word is computed.

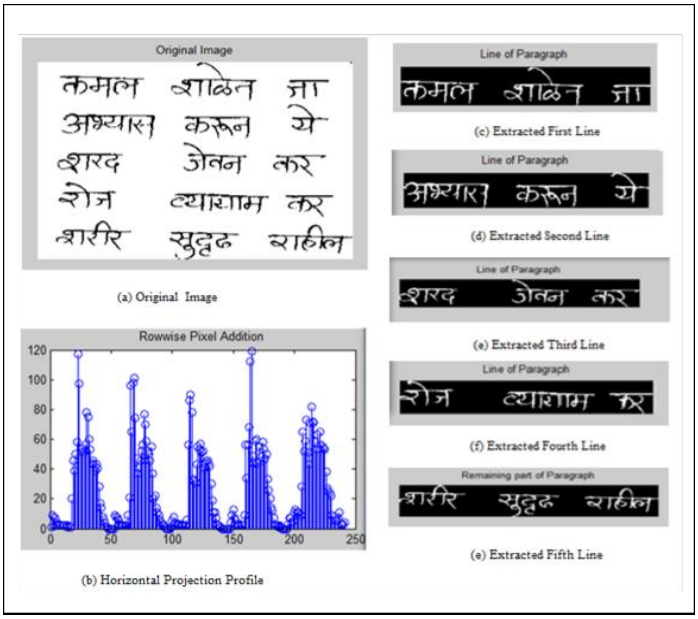

Fig: Segmentation Using OCR Technique.

Feature Extraction

This step describes the various features selected by us for classification of the selected characters. Extrac-tion of good features is the main key to correctly recognize an un-known character. A good feature set contains discriminating infor-mation, which can distinguish one object from other objects. It must also be as robust as possible in order to prevent

generating differ-ent feature codes for the objects in the same class.

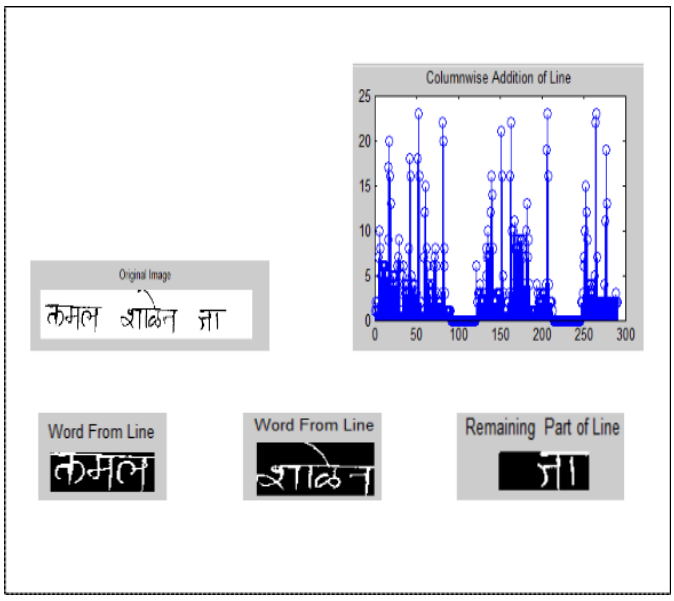

Fig: Feature Extraction In OCR Technique

Text analysis: concepts and technique:

i. **Before Information extraction (IE):** It identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process usually called pattern matching, typically based on regular expressions. The most popular form of IE is named entity recognition (NER). NER seeks to locate and classifyatomic elements in text into predefined categories (usually matching preestablished ontologies). NER techniques extract features such as the names of persons, organizations, locations, temporal or spatial expressions, quantities, monetary values, stock values, percentages, gene or protein names, etc. These are several tools relevant for this task: Apache OpenNLP, Stanford Named Entity Recognizer LingPipe .

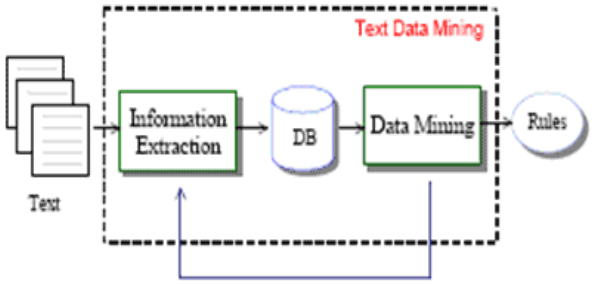

Fig: Overview Of Text Mining Framework.

ii. **K-means Clustering Algorithm:** Clustering is the classification of objects into different groups, partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait-often according to some defined distance measure. Once cluster are obtained then each cluster can be examined for other outcomes such as classification. The goal is to achieve high availability, scalability or both.

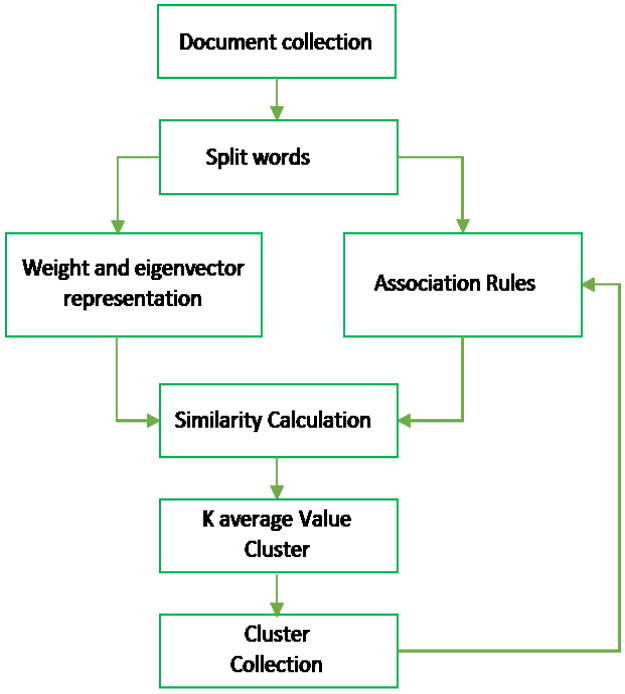

Fig: Document Clustering

Algorithm and flowchart

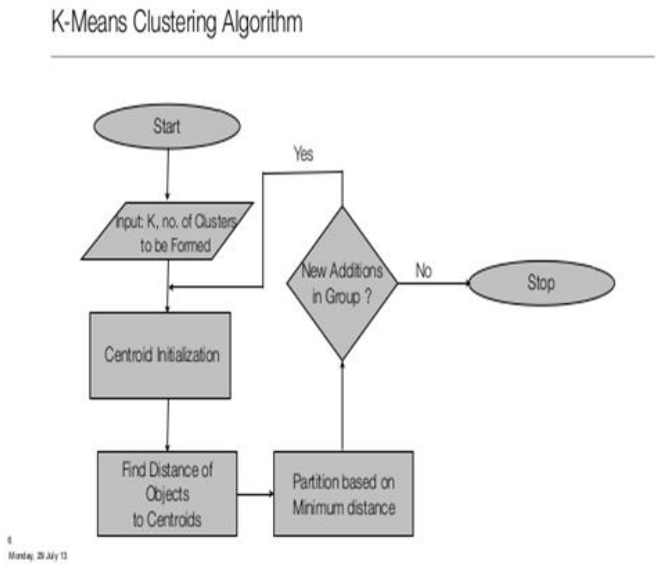

Fig: Flowchart Of K-Clustering Algorithm

Pseudo-code for K Means Clustering

```
Loop through K times
    current centroid = Randomly generate values for each attribute
Done = False
All instances cluster = none
WHILE not Done
    Total distance = 0
    Done = true
    For each instance
        instance's previous cluster = instance's cluster
        measure euclidean distance to each centroid
        find smallest distance and assign instance to that cluster
        if new cluster != previous cluster
                Done=False
        add smallest distance to total distance
    Report total distance
    For each cluster
        loop through attributes
                loop through instances assigned to cluster
                        update totals
                calculate average for attribute for cluster – producing new centroid
END While
```

Advantages

- Easy to compute.
- Some basic metric to extract the most descriptive terms in a document.
- It does not capture the position in text, semantics, co-occurances in different documents.
- Cannot capture semantics(e.g:as compared to word embeddings).

Conclusion:

The computing world has a lot to gain or benefits from various AI approaches. Their ability to learn by example makes them very flexible and powerful. Furthermore there is no need to devise an algorithm in order to perform a specific task i.e. there is no need to understand the internal mechanisms of that task. They are also very well suited for real time systems because of their fast response and computational times which are due to their parallel architecture. The goal of artificial intelligence is to create computers whose intelligence equals or surpasses humans. The proposed methodology will help in information retrieval in the field of scanned documents. The future work will be implementing the use of dictionary words to improve the performance of OCR system.

References

[1] Russell G and Perrone M (2012) International Workshop on Frontiers in Handwriting Recognition © IEEE

[2] Zadeh L.A. The concept of a linguistic variable and its application to approximate reasoning.

[3] P. Cowling, S. Remde, P. Hartley, W. Stewart, J. Stock-Brooks, T. Woolley(2010), "C-Link Concept Linkage in Knowledge Repositories". AAAI Spring Symposium Series.

[4] C-Link (2015): http://www.conceptlinkage.org

[5] Y. Hassan-Montero, and V Herrero-Solana (2016). "Improving Tag-Clouds as Visual Information Retrieval Interfaces", I International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2016.

[6] Wordle (2014): http://www.wordle.net.

[7] Xindong Wu, Senior Member, IEEE "Data Mining: An AI Perspective" vol.4 no 2 (2004)

[8] Satvika Khanna et al. "Expert Systems Advances in Education" NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, 19-20 March 2010

[9] Kaijun Xu." Dynamic neuro-fuzzy control design for civil aviation aircraft in intelligent landing system. Dept. of Air Navig. Civil Aviation Flight Univ. of China 2011.

[10] Eike.F Anderson.,"Playing smart artificial intelligence in computer games" The National Centre for Computer Animation (NCCA) Bournemouth University UK.

[11] K.R. Chaudhary "Goals, Roots and Sub-fields of Artificial Intelligence. MBM Engineering College, Jodhpur, India 2012

[12] Text Analytics: the convergence of Big Data and Artificial Intelligence(2016)-Universal Autonoma de Madrid and Instituto de Ingeniería del Conocimiento, Madrid, Spain 2ZED Worldwide, Madrid, Spain.

[13] Information Retrieval Using Artificial Intelligence & Fuzzy Logic For Documents Through OCR(2015) – PandeyM.K And Nandan Singh.

[14] Google – Word2vec (2016): http://arxiv.org/pdf.