INTERNATIONAL JOURNAL OF
RESEARCH IN COMPUTER
APPLICATIONS AND ROBOTICS
ISSN 2320-7345

COLOR & GREY SCALE IMAGE AND TEXT IDENTIFICATION USING PATTERN RECOGNITION.

T. SHANTHA KUMAR

Assistant .Professor, Department of Computer Science, Alpha Arts & Science College, Chennai.

Abstract: - Visual analytics is a dynamically emerging multidisciplinary research area which can be effectively used in the field of Visual Information System Management. Data collection and Data Management methodology plays a vital role in any scientific visual information system. The scientific literature is conveyed visually using plots, photographs, illustrations, diagrams, and tables. In existing system the pattern and visual information are not analyzed and classified. In proposed system, text and image processing is done by Text cue extraction and region extraction using vector construction. To facilitate research work on scientific figures in publication database (viziometrics). Utilize NLP and image processing techniques for inference from scientific figure. Combine text and visual cues for inferring details of research reported in a research article. Develop classification model for categorizing research articles based on inference from scientific figure.

INDEX TERM: - Text Summarization, Visual analysis, NLP, Viziometrics.

INTRODUCTION

Visual analytics is a dynamically emerging multidisciplinary research area which can be effectively used in the field of Visual Information System Management. Visual is the most elevated data transfer capacity data channel into the human mind and people are known to better hold data introduced outwardly. Most of the scientific literature contains vast of alpha-numeric data. Data collection and Data Management methodology plays a key role in any scientific visual information system. It is a key requirement to search and find information on the efficiency of the approach.

Past few years, Scientific Scholarly literature is published at a high frequency. Therefore, it is essential to develop framework to search and navigate the paper to identify various visual elements like plots, photographs, illustrations, diagrams, and tables.

MOTIVATION OF THE PROJECT

The main aim of the project is the existing method to processing the scientific literature is based on text processing in this novel method text and image based classification of scientific article.

LITERATURE SURVEY

Jevin et al. [1] proposed scholarly literature necessitate intelligent algorithms for search and navigation. Eigen factor recommends a citation based method for improving scholarly navigation. The hierarchical structure of scientific knowledge possible multiple scales of relevance for different users. A number of document collections and portals are already deploying basic recommendation designs that aim toward this goal.

PageRank approach on article-level citation graphics is the random walker on the graph will move inexorably backwards in time, and as a result will over-weight older papers.

INFORMATION EXTRACTION FROM FIGURES

Sagnik et al. [2] proposed modules an extractor for figures and associated metadata (Figure captions and mentions) from PDF documents. A Search engine on the extracted figures and metadata, Image processing module for automated data extraction from the figures, and a natural language processing module to understand the semantics of the figure. A semiautomatic system for data extraction from figures which is integrated with search engine to improve user experience.

IMAGE AND TEXT CORRELATION

Youtian et al. [3] proposed a set of documents, including an image and a textual description in the form of keywords, sentences or paragraphs. Image generally consists of a few visual patches. The correlated textual consists of meaningful keywords, and each keyword can be considered as a concept label of the visual patches. The label of visual patch under the supervision at the granularity of images and textual documents. Multi-Instance Multi-Label Learning is a learning paradigm is simultaneously represented by a bag of instances and associated with a set of
class labelsdistributionfromimagestothemultiplesubregionswithinthem.Animagecan is described using a small vocabulary of blobs, and computed the posterior probability of a word that corresponds to the given blobs of an image.

ANALYSIS VISUAL INFORMATION

PoShenetal. [4] Proposed Scientific results are communicated visually in the literature through diagrams, visualizations, and photographs. Information in the scientific literature is conveyed visually using plots, photographs, illustrations, diagrams and tables.

The practices for scientific communication. Image processing pipeline that classifies scientific figures into different categories to search interface that uses these classified images as the primary unit for exploring scholarly content.

DOCUMENT SUMMARIZATION

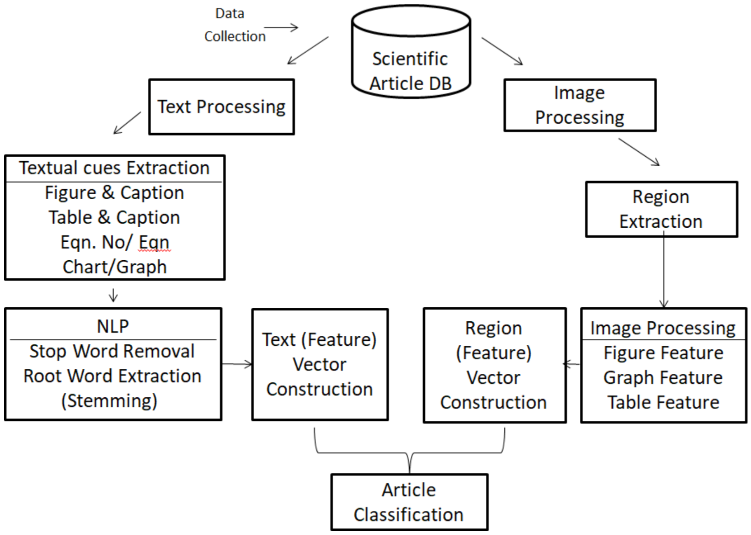
As the information resources are increasing tremendously, readers are overloaded with loads of information.

MOTIVATION OF THE PROJECT

The main aim of the project is the existing method to processing the scientific literature is based on text processing in this novel method text and image based classification of scientific article.

SYSTEM ARCHITECTURE DIAGRAM

The architecture diagram of article classification System is shown in Figure 3.1 illustrates the process is the Existing method to processing the scientific literature is based on text processing in this method text and image based classification of scientific article. In this system takes as input scholarly documents in PDF form2 .There are two process texts processing and image processing.



Text summarization is a process of extracting or collecting important information from original text and presents that information in the form of summary. Text summarization has become the necessity of many applications for example search engine, business analysis, market review. Summarization helps to gain required information in less time. The problem is termed as InformationOverload.Textsummarizationaddressesthisproblembyproducing the summary of related documents. Text summarization is one of the typical tasks of text mining.Is the process of reducing a text document with a computer program in order to create a summary that retains important points of the original document Extractive Summarization: These methods rely on extracting several parts, such as phrases and sentences, from a piece of text and stack them together to create a summary. Therefore, identifying the right sentences for summarization is of utmost importance in an extractive method.

Abstractive text summarization involves generating entirely new phrases and sentences to capture the meaning of the source document. This is a more challenging approach, but is also the approach ultimately used by humans. Classical methods operate by selecting and compressing content from the source document.

Extractive text summarization done by picking up the most important sentences from the original text in the way that forms the final summary. Extractive techniques generally generate summaries through 3 phases or it essentially based on them. These phases are pre-processing step, processing step and generation step

Pre-processing step: the representation space dimensionality of the original text is reduced to involve a new structure representation. It usually includes: a. Stop-word elimination: Common words without semantics that do not collect information relevant to the task (for example,” the”,” a”,” an”,” in”) are eliminated. b. Stemming: Acquire the stem of each word by bringing the word to its base form.
It uses an algorithm with the help of features generated in the preprocessing step to convert the text structure to the summary structure. In which, the sentences are scored.

ABSTRACTIVE SUMMARIZATION

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage. Abstractive summarization is the technique of generating a summary of a text from its main ideas, not by copying verbatim most salient sentences from text.
This is an important and challenge task in natural language processing .to abstractive text summarization based on discourse rules, syntactic

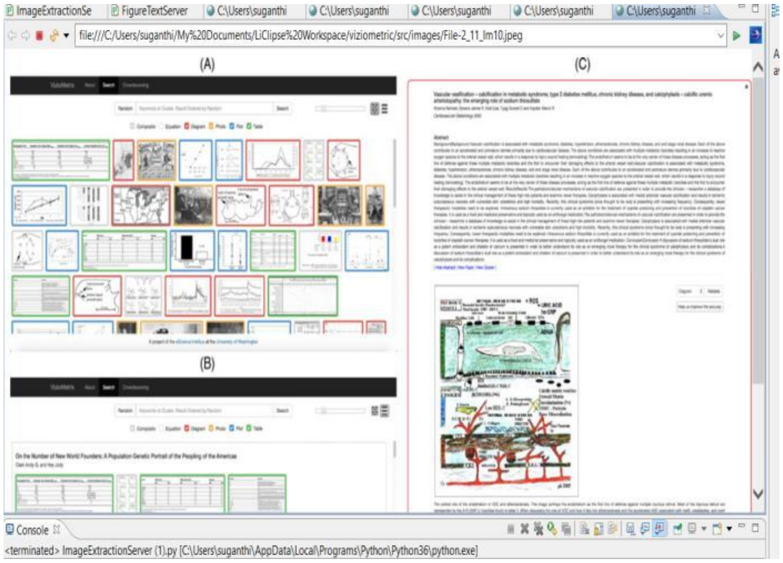
Constraints and word graph. Discourse rules and syntactic constraints are used in the process of generating sentences from keywords. Word graph is used in the sentence combination process to represent word relations in the text and to combine several sentences into one.

Experimental results show that our approach is promising in solving the abstractive summarization task.

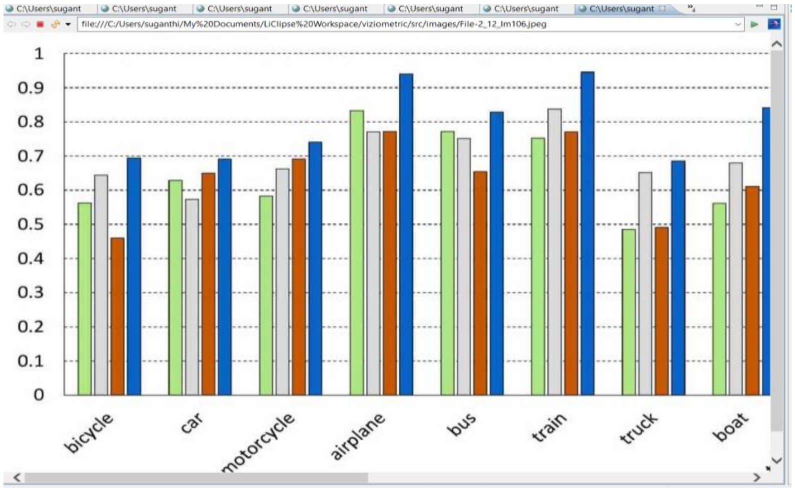
NATURAL LANGUAGE PROCESSING

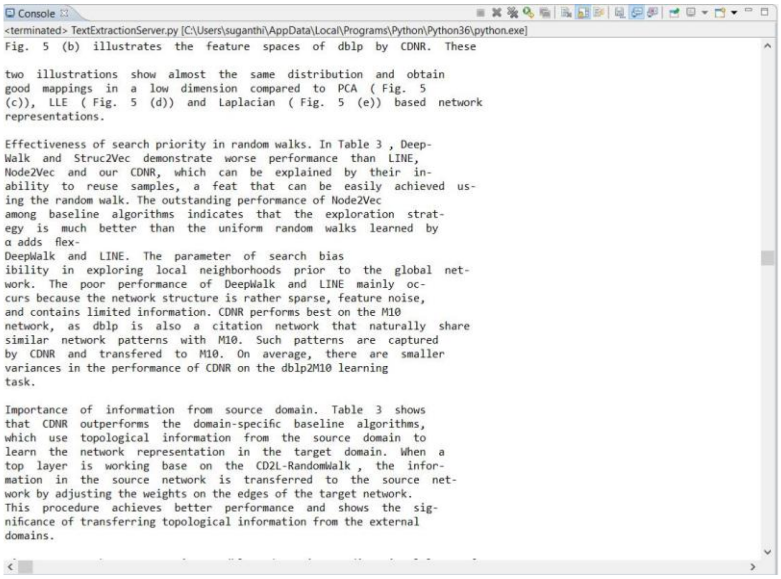
NLP can be used to interpret free text and make it analyzable. There is a tremendous amount of information stored in free text files, like patients' medical records, for example. Current approaches to NLP are based on deep learning, a type of AI that examines and uses patterns in data to improve a program's understanding. Deep learning models require massive amounts of labeled data to train on and identify relevant correlations, and assembling this kind of big data set is one of the main hurdles to NLP currently.

This involves chunking the text in the PDF into blocks, then identifying which blocks of text are captions, body text, or part of a figure. This step also attempts to identify regions containing graphical components. The output is a number of bounding boxes labeled as body text, image text, caption text, or graphic region.



```
<terminated> ImageExtractionServer (1).py [C:\Users\suganthi\AppData\Local\Programs\Python\Python36\python.exe]
```





```
<terminated> TextExtractionServer.py [C:\Users\suganthi\AppData\Local\Programs\Python\Python36\python.exe]
Fig. 5 (b) illustrates the feature spaces of dblp by CDNR. These

two illustrations show almost the same distribution and obtain
good mappings in a low dimension compared to PCA ( Fig. 5
(c)), LLE ( Fig. 5 (d)) and Laplacian ( Fig. 5 (e)) based network
representations.

Effectiveness of search priority in random walks. In Table 3 , Deep-
Walk and Struc2Vec demonstrate worse performance than LINE,
Node2Vec and our CDNR, which can be explained by their in-
ability to reuse samples, a feat that can be easily achieved us-
ing the random walk. The outstanding performance of Node2Vec
among baseline algorithms indicates that the exploration strat-
egy is much better than the uniform random walks learned by
α adds flex-
DeepWalk and LINE. The parameter of search bias
ibility in exploring local neighborhoods prior to the global net-
work. The poor performance of DeepWalk and LINE mainly oc-
curs because the network structure is rather sparse, feature noise,
and contains limited information. CDNR performs best on the M10
network, as dblp is also a citation network that naturally share
similar network patterns with M10. Such patterns are captured
by CDNR and transfered to M10. On average, there are smaller
variances in the performance of CDNR on the dblp2M10 learning
task.

Importance of information from source domain. Table 3 shows
that CDNR outperforms the domain-specific baseline algorithms,
which use topological information from the source domain to
learn the network representation in the target domain. When a
top layer is working base on the CD2L-RandomWalk , the infor-
mation in the source network is transferred to the source net-
work by adjusting the weights on the edges of the target network.
This procedure achieves better performance and shows the sig-
nificance of transferring topological information from the external
domains.
```

CONCLUSION AND FUTUREWORK

In this project to facilitate research on scientific figures, an area we call viziometrics. It extends prior work in bibliometrics and scientometrics but focuses on the role of visual information encodings. Developed a figure

Processing pipeline that automatically classifies figures into equations, diagrams, plots, photos, and tables. To facilitate further research on these visual objects, both the code and the data open for other researcher's to explore.
By integrating the figure-type labels and article metadata, analyzed the patterns across journals, overtime, and relationships to impact. To found that the role of the five figure types can vary widely. For instance, clinical papers tend to have higher photo density and computational papers tend to have higher diagram and plot density. High-impact papers tend to have more diagrams per page and a higher proportion of diagrams relative to other figure types.

BIBLIOGRAPHY

[1] H. Dave and S. Jaswal, "Multiple text document summarization system using hybrid summarization technique," in 2015 1st International Conference on Next Generation Computing Technologies (NGCT). IEEE, 2015, pp. 804–808.

[2] Po-Shen Lee , Jevin D. West, and Bill Howe "Viziometrics Analyzing Visual Information in the Scientific Literature "in the journal of IEEE Transaction on Big Data ,vol.4, No. 1, January-March 2018.

[3] S. Ray Choudhury and C. L. Giles, "An Architecture for information extraction from figures in digital libraries," in the Proceeding of 24th International Conference World Wide Web Companion, 2015.

[4] Youtian Du , Hang Wang, Yunbo Cui, and Xin Huang "Fundamental Visual Concept Learning From Correlated Images and Text" in the journal of IEEE Transaction on Image Processing, vol. 28, No. 7, July 2019.

[5] J. D. West, I. Wesley-Smith, and C. T. Bergstrom, "A Recommendation system based on hierarchical clustering of an article level citation network," in the journal of IEEE Transaction. Big Data, vol. 2, no. 2, pp. 113–123, July 2016.