Shodhshauryam, International Scientific Refereed Research Journal
© 2020 SHISRRJ | Volume 3 | Issue 4 | ISSN : 2581- 6306

Graph Based Framework for Visual Information Analysis in Scientific Documents & Scholarly Article using Graph Based Method

R. Mahalakshmi[1], K.Valarmathi[2]

[1]P.G Student, Department of CSE, Panimalar Institute of Technology, Chennai, Tamil Nadu, India

[2]Professor, Department of CSE, Panimalar Institute of Technology, Chennai, Tamil Nadu, India

ABSTRACT

Scientific results are communicated visually in the literature through diagrams, visualizations, and photographs and hence it is difficult to get the information from scientific literature quickly. The visualizations in the scientific literature to enhance search services, detect plagiarism, and study biblio metrics. An immediate problem is the ubiquitous use of multi-part figures: single images with multiple embedded sub-visualizations. The information content of the scientific literature is largely represented visually in the figures — charts, diagrams, tables, photographs, etc. Image processing is spreading in various fields. Image processing is a method which is commonly used to improve raw images which are received from various resources.

Keywords : Graph Based Framework, Edge Detection.

I. INTRODUCTION

Visual analytics is a dynamically emerging multidisciplinary research area that can be effectively used in the field of Visual Information System Management. Itis a business knowledge produce that works with Visual Data Center arrangement by gathering and breaking down both static and dynamic information to give merged, intelligent and visual diagrams and charts.

Visual Data Analytics is an incredible and significant instrument intended to empower its clients to envision the previous history, comprehend the present position and get ready for future patterns as far as space, power, and cooling. By coordinating with the business driving business insight device, Microsoft Power BI, a client can share any of the outlines and charts with different clients by using the security get to rights highlights.

The common outlines and diagrams can be seen inside a Web Browser so there is no necessity for other programming establishments

II. RELATED WORK

Visual analytics is a dynamically emerging multidisciplinary research area that can be effectively used in the field of Visual Information System Management.

It's a business knowledge produce that works with Visual Data Center arrangement by gathering and breaking down both static and dynamic information to give merged, intelligent and visual diagrams and charts. Visual Data Analytics is an incredible and significant instrument intended to empower its clients to envision the previous history, comprehend the present position and get ready for future patterns as far as space, power, and cooling. By coordinating with the business driving business insight device, Microsoft Power BI, a client can share any of the outlines and charts with different clients by using the security get to rights highlights. The common outlines and diagrams can be seen inside a Web Browser so there is no necessity for other programming establishments.

Bibliometrics was first coined by Wyndham Hulme in 1922 during his lecture at the University of Cambridge. Hulme identified the application of science and technology in the world of counting documents (Hulme 1923). Hulme's work on UK patents was the noteworthy literature in the initial decade of bibliometrics.

This is the time when bibliometrics was much explored to get transformed into Scientometrics. Golem(1943,1944) next used the term bibliometrics in his work on college libraries. Raising (1962) used the term couple of decades later in a critical essay on citation studies. Probably this was the first literature on the analysis of research citations. There is enough research work carried out since then on citation studies. Early days of statistical analysis only concentrated on citation counts and similar quantifying parameters.

Indicators like Journal Impact Factor (JIF) were also introduced which received wide acclaim over the bibliometric community.

Since bibliometric research was ruled by the library and information science community, the application of the statistical method was the modern means of research in bibliometrics. The h-index (Hirsch,2005) metric was introduced only in the current century and yet it is a widely accepted metric based on quantification of citations to the research article

As machine learning research was well established by the end of the 20th century, with more mature tools introduced to handle large data, there emerged a new research gap to apply statistical machine learning methods to derive new finding In bibliometrics research, As time progressed, the mere application of statistical techniques was not sufficed; this was mainly due to the lack of credibility or nonuniform research patterns studied across the world.

III. LITERATURE SURVEY

Rashmi et al.[7] Edge detection is basically, a method of segmenting an image into regions of discontinuity. Edge detection plays an important role in digital image processing. To compare different edge detection techniques for effective performance results concerning complex images. Edges characterize boundaries and are therefore a problem of fundamental importance in image processing. Many edge detection techniques have been developed for extracting edges from digital images. The purpose of edge detection is significantly reducing the amount of data in an image and preserves the structural properties for further image processing. Edge detection makes use of differential operators to detect changes in the gradients of the grey levels. The major classification of the edge detection technique or algorithm. Classical Operator, Laplacian of Gaussian, Gaussian. The Zero crossing operators perform the deduction of false zero crossings of an image and perform complex computation.

DISMANTLING COMPOSITE VISUALIZATIONS IN THE SCIENTIFIC LITERATURE

Po-Shen Lee et al. [8] The visualizations in the scientific literature to enhance search services, detect plagiarism and study bibliometrics. An immediate problem is the ubiquitous use of multi-part figures: single images with multiple embedded subvisualizations. a set of photographs of electrophoresis gels (B), an accumulation of a specific type of cells represented as a bar chart (C), and an alternative visualization of molecular sequences. Basic image segmentation techniques are inapplicable; they cannot distinguish between meaningful sub-figures and auxiliary fragments such as labels, annotations, legends, ticks, titles, etc.

IMAGE PREPROCESSING

Deepa et al.[15] Pre-processing is a common name for operations with images at the lowest level of abstraction - both input and output are intensity images. The aim of pre-processing is an improvement of the image data that suppresses unwilling distortions or enhances some image features important for further processing, although geometric transformations of images (e.g. rotation, scaling, translation) are classified among pre-processing methods here since similar techniques are used. Image preprocessing methods are classified into four categories according to the size of the pixel neighborhood that is used for the calculation of new pixel brightness. Image preprocessing methods use considerable redundancy in images. Neighboring pixels corresponding to one object in real images have essentially the same or similar brightness value, so if a distorted pixel can be picked out from the image, it can usually be restored as an average value of neighboring pixels.

AUTOMATIC UNSUPERVISED SHAPE RECOGNITION TECHNIQUE

Tudor Barbu et al.[11] The shape recognition of an introduction image shape analysis domain, a shape feature extraction technique using moment-based measures are invariant to geometric transforms. Then, an automatic unsupervised feature vector classification approach is proposed. It is based on a sequence of hierarchical agglomerative region-growing clustering algorithms and a measure based on cluster validation indexes.

The results of this provided recognition technique can be applied 9 successfully in important domains, such as object recognition, shape-based image content indexing, and retrieval. The most important region-based recognition techniques are those based on the Angular Radial Transform Descriptor (ARTD), Zernike moments and Legendre moments. Contour-based shape recognition includes techniques based on Fourier Descriptors, Curvature Scale Space Descriptors (CSSD), Contour Trees, Reeb Graphs and invariant moments.

IMAGE SPLIT-AND-MERGE METHOD

Yi Xiao et al. [12] Split-and-merge algorithm is one of the approaches for smoothing contours of an image. As the vertices of polygonal approximation are restricted to a subset of the original curve points, the Method can be used to compress the data. The main function in the split-and-merge algorithm is the collinearity test that checks if points along a boundary are collinear concerning a straight line. Collinearity is usually determined by the maximum perpendicular distance from a point on the boundary portion to the straight line. Thresholding Algorithm to quantization error, a straight line is in a zigzag Shape.

The aim is to replace the zigzag digital arc of a straight line by its two terminal points. That is, when the collinear test in the split-and merge algorithm is performed, the maximum perpendicular distance from any point on the digital line to the straight line joining the two ends of the digital line is less than the tolerance.

To find the maximum distance value mentioned above, the character of the chain code of a straight line has to be known. It is well known that the chain code of a straight line must satisfy three criteria. The polygonal approximation is a useful tool for data reduction when representing digital contours. Its accuracy depends on the value of tolerance. When the selection of tolerance is adaptive, the split-and-merge process can provide a set of vertices with high information content about the contour. That is, the quantization error is attenuated and the original shape is preserved.

SVM CLASSIFICATION ACCURACY USING A HIERARCHICAL APPROACH

Begüm Demir et al. [14]SVM classification with a hierarchical approach to increase SVM classification accuracy as well as reduce the computational load of SVM testing. Support vectors are obtained by applying SVM training to the entire original training data. For classification, multi-level two-dimensional wavelet decomposition is applied to each hyperspectral image band and low spatial frequency components of each level are used for hierarchical classification.

Initially, conventional SVM classification is carried out in the highest hierarchical level (lowest resolution) using all support vectors and a one-to-one multiclass classification strategy, so that all pixels in the lowest resolution are classified. In the subsequent levels (higher resolutions) pixels are classified using the class information of the corresponding neighbor pixels of the upper level.

EDGE DETECTION

Edge detection is important in the image preprocessing, and the purpose is to detect the image in the brightness of the object edge. The edge direction as shown in fig 3.2 of the image is different, the amplitude of the change is also different, the horizontal pixel value changes very slowly, and in the vertical direction of the transformation is more intense, so the first or second-order differential operator on the image of the object edge detection.

Edge is the most obvious change of gray level on the image. Edge detection uses this feature to locate the edge pixels of each pixel of the image by differential or two order differentiation. According to the gray values of adjacent areas, the edge types can be divided into three steps: ladder shape, pulse shape, and roof shape.

For the ladder shape, the image edge points correspond to the peak values of the first-order differential image and the zero crossings of the two order differential image; For pulse-like and roof like edges, the edge points correspond to the zero crossings of the first derivative and the peak of the two derivatives

IMAGE COMPRESSION

Image compression signifies compression of the records among the digital images. Image compression eliminates duplication of the data so that it will be stored and transmitted effectively. Image compression might be lossy and lossless. In lossless compression, before and after compression the quality of data remains consistent.

In lossy compression, the quality of data decreases after applying the compression techniques. Lossless compression is mostly used for medical imaging, technical drawing contents, and archival purposes, etc. Lossy approaches are used in those environments in which minor loss of quality is acceptable to accomplish a considerable reduction in bit rate. The most widespread technique for compression is JPEG which compresses full color or grayscale images. JPEG uses discrete cosine transforms technique for compression. There is another technique for compression known as a Wavelet transform. Through wavelet, data is divided into different frequency components and then the further study is done for each component. Wavelets have advantages over traditional Fourier approaches in examining physical circumstances.

IV. CONCLUSION

The proposed system is a complete architecture for the classification of scholarly documents and identifying the edges in the given graph. The architecture contains multiple modules to perform defined tasks, which include Graph Identification & Edge Identification of the Given Input Images. The report a novel method for the extraction of information from PDF documents, a simple but effective classifier for extracted information and identifying the graph-based information from the given input files. The method is based on the hierarchical structure of scientific knowledge, allowing for different scales of influence.

V. FUTURE WORK

The proposed system involves the developing metadata formats for different types of information and developing algorithms for fully automatic data extraction from different types of contents from the given input files. In the future, plan to use the extracted data and metadata to create a natural language processing and identifying the various edges from the given input files. Article classification is discovering some emerging trends in the domain of computer science and engineering and the most recent research topic of article classification is adaptive control

VI. REFERENCES

[1] 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* Proceeding of IEEE Trans. Big Data, vol. 2, no. 2, pp. 113–123, Apr.-Jun. 2016.

[2] Po-Shen Lee, Jevin D. West, and Bill Howe “Viziometrics Analyzing VisualInformationintheScientificLiterature”*in the* journal ofIEEETransactiononBigData, 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, pp. 667–672.

[4] L. Bornmann and R. Mutz, “Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references,” J. Assoc. Inform. Sci. Technol., vol. 66, pp. 2215–2222, 2015.

[5] GraphIE: A Graph-Based Framework for Information Extraction-Yujie Qian1, Enrico Santus1, Zhijing Jin2, Jiang Guo1, and Regina Barzilay 2019, pages 751–761.

[6] S. R. Choudhury, et al., “Figure metadata extraction from digital documents,” in *in the*Proceeding of 12th International Conference Document Anal. Recog., 2013, pp. 135–139.

[7] Rashmi, Mukesh Kumar, and RohiniSaxenaetal[1]“Algorithm And Technique On Various Edge Detection” *in the* Proceeding of International Journal (SIPIJ) Vol.4, No.3, June 2013, pages 65–75.

[8] Po-Shen Lee, and Bill Howe “Dismantling Composite Visualizations in the Scientific Literature”*in the* Proceeding of International conference,January 2016,pp 247-266.

[9] Dejian Yu, ZeshuiXu, Senior Member, IEEE, Yuhsuan Kao, and Chin-Teng Lin, Fellow “The Structure and Citation Landscape of IEEE Transactions On Fuzzy Systems”-*in the*Proceeding of IEEE Transactions On Fuzzy Systems, Vol. 26, No. 2, April 2018.

Cite this article as :

R. Mahalakshmi, K.Valarmathi, "Graph Based Framework for Visual Information Analysis in Scientific Documents & Scholarly Article using Graph Based Method", Shodhshauryam, International Scientific Refereed Research Journal (SHISRRJ), ISSN : 2581-6306, Volume 3 Issue 4, pp. 16-20, July-August 2020.
URL : http://shisrrj.com/SHISRRJ120321