Abstract

Information Retrieval (IR) from Tamil document images present in World Wide Web (WWW) has become a challenging problem today due to its rising popularity. Among the most valuable Web assets, categorizing web images and retrieval of information from the images on the Web is quite difficult. This paper proposes a simple and effective method to separate the document images from the available web image sources and to retrieve the information present in those web document images. This system works in two phases: In the first phase, it concentrates on Automatic Image categorization process over web images by employing a filtering technique to discriminate the document images from other images available in WWW. Filtering technique employed here captures the image information by intensity and frequency histograms to discriminate the web document images. As for information retrieval in the second phase, feature string generation technique has been used to generate feature strings for every word images by extracting its shape this generates a feature string for every word image by extracting its features relying on their statistical properties, such as lines, black and white disposition rates and outline features of characters, instead of recognizing the letters and assigning its ASCII value like OCR. This kind of information retrieval has been initiated over a list of web sites and experimental results are recorded.
References


**Index Terms**

Computer Science

Information Sciences

**Keywords**

Web Search  Information retrieval  Web Image categorization  Document Images