Short Write-up On research proposal on Internet Multimedia Search Engine for Information Retrieval in distributed Environment (DST No: DST/TSG/ICT/2009/27 Dated 03-09-2010)

This research project has proposed approaches to integrate both text-based and content-based retrieval system with a suitable indexing mechanism for indexing the features extracted. Both low-level and high-level features along with the HTML documents are fetched from WWW and stored in respective repository. As a result of voluminous content, there is a need for huge space to archive them. A novel tree based data structure is proposed to represent the features and HTML documents in a compact form. This reduces space requirement of crawling document along with the features considerably and also make the retrieval faster. In addition, a suitable coding scheme based on Colombo-rice coding is proposed for representing the low-level features in a compact form. The number of bits used for representing the low-level features is considerably reduced. The bins with zero values are not considered and a suitable similarity measure is proposed for calculating the similarity distance between the query and database features.

The proposed research has come up with a sequential pattern based tree structure for enabling query refinement for image retrieval. In addition, more numbers of low-level features and high-level features are combined during any stage of retrieval for refining the query to narrow the search space for improving the precision of retrieval. The entire query refinement is decomposed into small component with defined activities as protocol and specification. While MHCPH is used as a low-level feature the weighted HTML TAG values are used as a source for high-level features. The MHCPH is proposed by using both the properties of the HSV color space and human visual perception system. The bin values are updated based on the neighboring color bin with a suitable weight. On the other hand for high-level features, the HTML TAGS are divided into six groups and weighted based on their capability in describing the semantics of the HTML pages. Both of these features are combined and represented in the form weighted TAG tree for facilities query refinement at any stage of retrieval phase.

This research project has developed a suitable architecture specification for a deep web crawler with surface web crawler as well as an indexer for fetching a large number of documents from deep web using rules. The functional dependency of core and allied fields in the FORM is identified for generating rules using SVM classifier and it classifies them as most preferable, least preferable and mutually exclusive. The FORMs are filled with values from most preferable class for fetching a large number of documents. The extracted document is indexed for information retrieval applications. The architecture is extended to distributed crawler using web services. The proposed crawler fetches a large number of documents while using the values in most preferable class. This architecture has higher coverage rate and reduces fetching time. The retrieval performance is encouraging and achieves similar precision of retrieval as Google search engine system.

A research attempt is made to convert the findings to develop a real-time search engine system for tracking criminal activities in the WWW. The principle of event detection identifies interesting events from web pages and a new approach is proposed to identify the sentences with an interested event description. The "event triggered" terms, co-occurrence terms along with the sentences are extracted from web documents. The patterns of sentence are analyzed by POS tagging and clustered using decision tree to identify "event mentions". Set of rules are generated which specifies "event mention" patterns and are prioritized based on the importance. The event mention" patterns are assigned the weights based on the semantic relation between the terms in the sentence to identify the "event instance". There exists common features between the sentences and there is no clear boundary. This aspect is captured using suitable fuzzy rules and the impreciseness is captured. Fuzzy logic based Artificial Neural Network (ANFIS) is constructed for training the sentence patterns to learn each instance of the event. The proposed approach is finding the event patterns effectively and compared to some of the recently proposed similar approaches.

The events extracted from the sentences are combined with the images present in a web page. While the images are human faces, the emotions are extracted and combined for better understanding of the pages. The emotions are the emergent property of the human mind like consciousness. Emotions are emerging from the interaction among various core cognitive processes. Cognitive models are available for other core cognitive processes like problem solving, decision making, memory, reasoning, etc.,. Since it is not clear about the origin of emotion, through emotional response, it is possible to create cognitive models of emotion. One kind of physiological or emotional response to the particular event via face is the facial expression. Human face plays an important role in interpersonal communication, which gets more importance compared to other communication channels while conveying mixed mode of information. Facial expression provides more visual cues about internal mental states of human being, which is caused by muscle contraction to facial skin change the appearance of facial features, say facial components such as the eyebrows, nose, and mouth. As facial expressions are natural feedback to others, it can be used in a wide variety of computer applications in which the system can recognize the human internal emotional state from the visual cues and can react accordingly. By developing valid and reliable methodologies to measure facial behavior, it can be used as a natural interface in various applications such as human–computer interaction, computer surveillance, gaming, entertainment, teleconference, medical field, education, etc.,.

Dr. A Vadivel
PI & Associate Professor
Department of Computer Applications
National Institute of Technology, Trichy