International Journal on Recent Researches In Science, Engineering & Technology
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy.
It is an absolutely free (No processing charges, No publishing charges etc) Journal Indexed in DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com
Chief Editors 1 : Dr. M.Narayana Rao, Ph.D., Rtd. Professor, NIT, Trichy.
(Engg.&Technology division)
2 : Dr. N.Sandyarani, Ph.D., Professor,
Chennai based Engg.College, (Science division)

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105
Volume 2, Issue 2, February 2014
DIIF IF :1.46
SJIF IF: 1.329

Web Data Extraction Using Clustering Techniques

Venkata Kishore.Konki and SameerGogineni

Abstract: The number of users using the web is growing rapidly, Because of the enormous growth of information on web. Due to this reason it creates many problems while retrieving information and finding solution to this problem became the current research topic. Data Extraction is one of the process for information retrieval from various data ware house applications like e-commerce, and other storage data bases. The availability of a vast amount of information on the Web is due the lack of a central structure and freedom from a strict syntax. Most users might miss relevant information because they just view the top ten results. The aim of clustering is either to create groups of similar objects or create a hierarchy of such groups. We focus here mainly on document clustering, e.g. objects are texts, web pages, phrases, etc. Generally, clustering approaches could be classified in two broad categories: term-based Clustering and link-based clustering. The algorithm proposed in this paper for clustering the web documents is described is Fuzzy C-Means Algorithm. Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees (between 0 and 1) and vague or fuzzy boundaries between clusters. The algorithm used for clustering the web documents described in this paper is represented graphically in the following figure that shows the steps to obtain the cluster of web documents which was discussed in the proposed approach.