© UNIVERSAL RESEARCH REPORTS | REFEREED | PEER REVIEWED
ISSN : 2348 - 5612 | Volume : 05 , Issue : 04 | January – March 2018

A Review Paper: Categorizationof Web Pages

Ritu[1], Sapna Aggarwal[2]
Student,M.TECH(CSE) Jind Institute of Engineering & Technology, Jind,India[1]
Assistant Professor, (CSE) Jind Institute of Engineering and Technology, Jind, India[2]
mail id@gmail.com[1]
er.sapna.aggarwal@gmail.com[2]

Abstract
Contemporary web is comprised of trillions of pages and everydaytremendous amount of requests are made to put more web pages on the WWW. It has been difficult to manage information present on web than to create it. Web pagecategorization can be defined as an approach to categorize the web pages based on a setof predefined categories to manage large web content. Yahoo! and ODP are theexamples of web directories in which pages are categorized manually or semiautomatically, but it is a very time consuming task. There are many ways of categorizingweb pages using different techniques. An approach to categorize webpages automatically on the basis of characteristics of web pages using neural networkbased single discrete perceptron training algorithm which is extended by selecting webpage specific features to categorize web pages of predefined categories with highaccuracy. The idea is presented with the help of two specific and major categories of webpages chosen for categorization that are newspaper and education.

Keywords: *Categorization, Web*

1. Introduction

Internet is the source of enormous amount of information accessed by large number of people every day. Contemporary web is comprised of trillions of pages and everydaytremendous amount of requests are made to put more web pages on the WWW. It hasbeen difficult to manage information present on web than to create it. Web pagecategorization can be defined as an approach to categorize the web pages based on a setof predefined categories to manage large web content. Yahoo! and ODP are theexamples of web directories in which pages are categorized manually or semiautomatically, but it is a very time consuming task. There are many ways of categorizingweb pages using different techniques. An approach to categorize webpages automatically on the basis of characteristics of web pages using neural networkbased single discrete perceptron training algorithm which is extended by selecting webpage specific features to categorize web pages of predefined categories with highaccuracy.

ISSN : 2348-5612 © URR

The growing number of applications on the web leads to rapid increase in number of web pages. The data available on the web can be in the form of text, images, audio, video, graphics and many other forms. Web pages present on the web can be static or dynamic. The content of dynamic web pages keeps on changing time to time. Web is considered as a large repository of information which is accessed by millions of users' everyday through internet. The dynamic nature of web and large scale explosion of web pages may put a threat to efficient information retrieval tasks. Web can be considered as an information resource, therefore it is important to describe and organize the huge content present on the web in order to realize web's full potential. Thus web page categorization is an intellectual task, important and indeed essential for organizing and understanding web content for different applications, efficient information retrieval and other tasks related to web mining. Here we will discuss some facts about web page categorization including the types of web page categorization, need of web page categorization and various characteristics of web pages.

2. Literature Review

From the very beginning categorization was done manually by domain experts. Yahoo! [3] and ODP [4] are the examples of web directories which are developed manually. But with the rapid increase of web pages it became extremely difficult to categorize web pages manually. Therefore categorization began to be done semi automatically or automatically. There are a number of approaches which have been applied in the field of web page categorization including K-Nearest Neighbor approach [11], Bayesian probabilistic models [12], inductive rule learning, decision trees, neural networksand support vector machine. All the above mentioned approaches