# A survey on Web mining: Data mining approach

**Nilesh Magar [1] , Aniket Nagane[2]**
[1 2] MAEER's Arts, Commerce and Science College, Pune(India)
[1] nileshraghav@gmail.com, [2] aniket.nagane@gmail.com

## ABSTRACT:

**The Web has become very important and versatile tool for almost all application in day today life. Internet is a hug source of information. Unfortunately information from many sources is not available in the form which could be mined properly. Web mining deals with extracting the interesting patterns and developing useful abstractions from diversified sources. Paper deals with the comparative study of preprocessing tasks, applications and challenge of web mining.**

**Key words:** web mining, knowledge discovery

## I INTRODUCTION

Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late [2]. Internet can be considered as a huge semi structured database, presenting all the problems implicit in semi-structured data. Extracting the structure of every HTML document is a challenging issue given the absence of predefined standard and schema [3]. But an internet is continuously becoming a central part of social, cultural, political, educational, academic and commercial life as it contains a wide range of information and applications.

Web mining can be broadly defined as the discovery and analysis of useful information from the internet[4]. Data can be collected from various sources like the server side, client side, proxy servers, or it can be obtained from an organization's database. This data is mostly large and hyperlinked. Depending on the location of the source, the type of collected data differs. It also has extreme variation both in its content (e.g., text, image, audio, symbolic) and Meta information that might be available. This makes the techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are[4].

1) Unlabeled;

2) Distributed;

3) Heterogeneous (mixed media);

4) Semi structured;

5) Time varying;

6) High dimensional.

Web mining tasks can be summarized as follow:

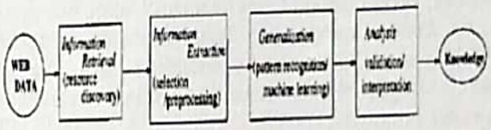
WEB DATA
Information Retrieval (resource discovery)
Information Extraction (selection /preprocessing)
Generalization (pattern recognition/ machine learning)
Analysis validation/ interpretation
Knowledge

Fig. 1. Web mining subtasks.

**Information Retrieval:** This process deal with the retrieval of relevant information from web data. The IR process mainly includes document representation, indexing, and searching for documents [4]. While retrieving information, methodologies used should ensure that non-relevant information should not extracted or very few amount of non relevant information should be retrieved.

**Information Selection/Extraction and Preprocessing:** once relevant information is retrieved from web data, important task is to extract knowledge and other required information without human interaction. The major methods of information retrieval involve writing wrappers (hand coding) which map the documents to some data model[4]

**Generalization:** Here in this phase pattern recognition and machine learning methods are used on the extracted data. A major problem while web mining is the labeling problem: hug amount of data is available on the web but it is unlabeled. Data mining techniques require proper labeled data as positive (yes) or negative (no) An

*ISBN NO : 978-93-84654-10-3*