

Chapter 1 : INTRODUCTION

Problem Definition: Web page classification, also known as Web page categorization, is the process of assigning a Web page to one or more predefined category labels or classes. Classification is traditionally posed as a supervised learning problem in which a set of labelled data is used to train a classifier which can be applied to label future examples.

As the number of pages on the web are permanently increasing, there arises a necessity to classify pages into categories to facilitate indexing or searching them. In the method proposed here, we use both textual and numerical information to find a suitable class of the web page based on its contents.

Classification plays a vital role in many information management and retrieval tasks. On the Web, classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific Web link analysis, to contextual advertising, and to analysis of the topical structure of the Web. Web page classification can also help improve the quality of Web search.^[3]

StumbleUpon is a user-curated web content discovery engine that recommends relevant, high quality pages and media to its users, based on their interests.^[5] While some pages *StumbleUpon* recommends, such as news articles or seasonal recipes, are only relevant for a short period of time, others maintain a timeless quality and can be recommended to users long after they are discovered. In other words, pages can either be classified as "ephemeral" (pages that lose their importance with time) or "evergreen" (web pages that are timeless due to the quality of information they carry). The ratings obtained from the user community gives us strong signals that a page may no longer be relevant - but what if this distinction could be made ahead of time? A high quality prediction of "ephemeral" or "evergreen" would greatly improve a recommendation systems.

Many people know evergreen content when they see it, but can an algorithm make the same determination without human intuition? Can a

machine - by its own intelligence - do this job of classifying these pages into appropriate categories so as to improve the recommendations of a recommendation engine. This would help in enhancing the user experience of a web search and he/she would be directed in his/her own interest path while browsing the Internet.

The *mission of our project* is to build a classifier which will evaluate a large set of URLs and label them as either evergreen or ephemeral. We have a collection of random URLs as our dataset. This includes a total of 10566 URLs out of which 7395 URLs are classified into Evergreen or Ephemeral and this forms our Training dataset. The remainder of 3171 URLs which forms our Test dataset, we would classify these into Evergreen or Ephemeral by adding intelligence to the machine learning algorithms with the help of our Training dataset.^[5] In other words we are training our classification model according to the training dataset and implementing it on the test dataset to predict the class of URLs in the test dataset.