RESEARCH ARTICLE OPEN ACCESS

Collective Behaviour Learning :A Concept For Filtering Web Pages

[1]G. Mercy Bai,[2]T. Selva Banupriya
[1]PG Student, [2] Lecturer
Department of Computer Science and Engineering
DMI College Of Engineering, Chennai-600123, India.

Abstract—
The rapid growth of the WWW poses unprecedented challenges for general purpose crawlers and search engines. The Former technique used to crawl web pages was FOCUS (Forum Crawler Under Supervision).This project presents a collective behavior learning algorithm for web crawling. The collective behavior learning algorithm crawl the web pages based on particular keyword. Discriminative learning extracts only the related URL of the particular keyword based on filtering. The goal of this project is to crawl relevant forum content from the web with minimal overhead. The unwanted URL is removed from the web pages and the web page crawling is reduced by using the collective behavior learning. The web pages must be extracted based on certain learning techniques and can be used to collect the unwanted URL'S.
Keywords— Uniform Resource Locator, World Wide Web, Support Vector Machine, Forum Crawler Under Supervision.

I. INTRODUCTION

Data mining is the process of analyzing data from different fields and summarizing it into useful information. Data mining software is one of the analytical tools for analyzing data. Users can analyze data from many different areas and summarizes the relationships identified. Data mining is the process of finding data or patterns in large relational databases for collecting different data. With the development of information retrieval websites have become a more and more important information medium for different organizations and people all over the world. The information is distributed on the web pages with some unwanted URL'S. The major aim of this project is to remove the unwanted URL'S. In this method for learning regular expression patterns of URLs that lead a crawler from an entry page to target pages. In the initial study, modularity maximization was employed to extract social URL. The superiority of this framework over other representative relational learning methods has been verified with social media data URL. The real framework, however, is not scalable to handle large sizes because the extracted social URL dimensions are rather dense. In social pages, a network of millions of URL is very common. With a huge number of URL extracted cannot even be held sometimes in memory, causing a serious computational problem. Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data mining works by analyzing data stored in data warehouses that are used to store that data that is being groped to analyze. The data may come from all parts of business, from the production to the management. Managers use data mining to decide upon marketing strategies for their own product generated.

A recent framework based on social URL'S dimensions is shown to be effective in addressing this heterogeneity. The framework shows a better way of URL classification: first, capture the latent affiliations of actors by extracting social dimensions based, and next, apply exact data mining techniques to classification based on the extracted dimensions. The collective behavior learning algorithm provides the basis for converting the wanted and unwanted URL. The web page is extracted based on certain URL extraction features and can be based on dimensional values. The unwanted URL'S are found based on certain features and can be extracted with social values. There are various kinds of valuable semantic information about real-world entities embedded in web pages and databases. Extracting and integrating the wanted information from the Web is of great significance. To extract the dimension of the web pages regularization is applied. Learning regular expression patterns of URLs that lead a crawler from an entry page to target pages. In the initial study, modularity maximization was employed to extract social URL. The superiority of this framework over other representative relational learning methods has been verified with social media data URL. The original framework, however, is not scalable to handle networks of sizes because the extracted social

URL dimensions are rather dense. In social media, millions of URL'S are very common. With a huge number dimensions cannot even, causing a serious computational problem.

II.COLLECTIVE BEHAVIOUR LEARNING ALGORITHM

The collective behavior learning algorithm provides the basis for converting the wanted and unwanted URL. The web page is extracted based on certain URL extraction features and can be based on dimensional values. The unwanted URL'S are found based on certain features and can be extracted with social values. There are various kinds of valuable semantic information about real-world entities embedded in web pages and databases. Extracting and integrating these entity information from the Web is of great significance. Regularization is applied to extract the dimension of the web pages.

STEPS IN COLLECTIVE BEHAVIOUR LEARNING

Input: network data, labels of some URL, number of social URL dimensions;
Output: labels of unwanted URL.
Steps:-1. Convert network into edge-centric view.
2. Perform edge clustering.
3. Construct social URL dimensions based on edge partition node belongs to one community as long as any of its neighboring edges is in that community.
4. Apply regularization to social URL dimensions.
5. Construct classifier based on social URL dimensions of labeled nodes.
6. Use the classifier to predict labels of unwanted ones based on their social dimensions.

III. MODULE DESCRIPTION

1. AUTHENTICATION:

In the authentication module user and admin login are used to provide information about the user of the web pages .the user and admin login are defined below

USER SIGN UP AND LOGIN:

In this module user can create account with the sites by filling details. The user can create account by using username and password. The username can consist of alphabet and special characters. The passwords consist of alphabet and numerals. The user will be authenticated only if they provide the correct username and password.

ADMIN LOGIN:

In admin login the admin has their username and password of their own and they authenticate the users who are entering into the system are authenticated or not. Admin provides a clear view of the system authentication. The admin has the right to create, delete and add pages and users. The admin can also
1) Create account
2) Delete account and
3) Modify account.

2. SOCIAL WEB PAGE EXTRACTION:

The latent social URL dimensions are extracted based on network topology to capture the potential affiliations of URL. These extracted social URL dimensions represent how each actor is involved in diverse affiliations. These social URL dimensions can be treated as features of actors for subsequent discriminative learning. Data is converted into features, important classifiers such as support vector machine and logistic regression can be used. Social dimensions extracted according to the data clustering, such as modularity maximization and probabilistic methods.

The web page is extracted based on certain URL extraction features and can be based on dimensional values. The unwanted URL'S are found based on certain features and can be extracted with certain values. There are various kinds of valuable semantic information about real-world entities embedded in web pages and databases.

3. DISCRIMINATIVE LEARNING:

Discriminative learning provides the way for extracting useful URL'S from unwanted URL'S. It is the discriminative model that provides efficient classification details. Discriminative learning technique is based on probability determination of probability values.

4. WEB PAGE CRAWLING:

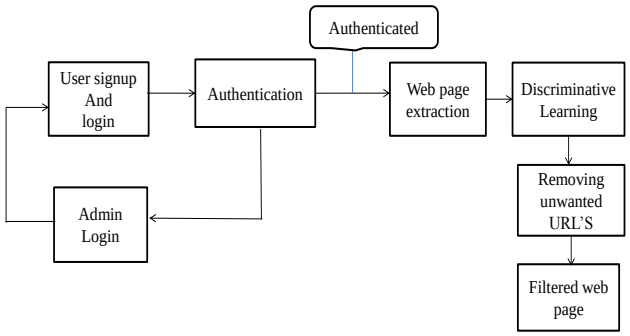
A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a systematic manner . This is called Web crawling or spidering. Search engines, use spidering as a means of which provides up to date information. Web crawlers are used to create a copy of all the visited pages for later processing by search engines , that will load the downloaded pages to provide fast searches in the websites. Web crawlers can be efficiently used used for automating maintenance of tasks on a Web site. Websites are checking links or validating HTML code. Crawlers are used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam).

IV.FLOW DIAGRAM

The flow diagram represents the stages involved in extracting the web pages and providing the login details and the authentication details. Each web page is extracted based on some features and the unwanted URL removal is done. The web pages with unwanted URL'S are removed and only the limited URL'S are generated. By removing the unwanted

pages each and every pages can be free for crawling and can be used with specified time limit. Discrimnative learning provides an efficient way for classifying the wanted and the unwanted URL'S. The unwanted URL'S are found based on certain features and can be extracted with social values. There are various kinds of valuable semantic information about real-world entities embedded in web pages and databases.

Fig : Flow diagram for web page filtering



V.CONCLUSION AND FUTURE ENHANCEMENT

The collective behavior learning algorithm provides an efficient way for extracting the web pages based on wanted and unwanted URL. The web page is extracted based on certain URL extraction features and can be based on dimensional values. The unwanted URL'S are found based on certain features and can be extracted with social values. There are various kinds of valuable semantic information about real-world entities embedded in web pages and databases. Integrating and extracting the entity information from the Web is of great significance.

The social pages will consist of advertisements and different violence and censored URL. This work provides a clear view of extracting wanted and unwanted URL. Social web page extraction is done and the unwanted URL'S in the web pages are removed. This is shown by the screenshots and each value is filtered by using certain discriminative leaning techniques. The main contribution is the web page classification based on different attributes. Social dimensions are extracted to represent the potential affiliations of actors before discriminative learning occurs. The approaches to extract social dimensions suffer from scalability, it is imperative to address the scalability issue. Propose an edge-centric clustering scheme to extract social dimensions and a scalable k-means variant to handle edge clustering. Each edge is treated as one data instance, and corresponding features are the connected nodes.

In our future work filtering the URL'S automatically by using specified tools will be performed. The advance work will limit the unwanted URL'S by applying certain dimension in high capacity values.

References

[1] Zhouyao Chen, Ou Wu, Mingliang Zhu and Weiming Hu " A Novel Web Page Filtering System by Combining Texts and Images", IEEE Transaction 2010.

[2] Xunxun Chen,Wei Wang, Dapeng Man and Sichang Xuan" A Webpage Deletion Algorithm Based on Hierarchical Filtering", IEEE Transaction 2010

[3] Marco CovaDavide, Canali and Giovanni Vigna "Prophiler: A fast filter for the large-scale detection of malicious" **,** SPRINGER Journal 2011.

[4] K.S.Kuppusamy and G.Aghila "A personalized web page content filtering model based on segmentation" **,**IEEE Journal 2011.

[5] Zhu He ,Xi Li andWeiming Hu "A boosted semi-supervised learning framework for web page filtering", IEEE Transaction 2010.

[6] Maria , Gomez Hidalgo, Francisco Carrero Garcia, and Enrique Puertas Sanz" Web Filtering Using Text Classification",Springer 2011.

[7] Tien Dung,Do,Kuiyu Chang Siu and Cheung Hui Tien"Web Mining for Cyber Monitoring and Filtering" ,IEEE Transaction 2011.