Available online at www.jpsscientificpublications.com
Volume – 1; Issue - 2; Year – 2015; Page: 135 – 139

Indo – Asian Journal of Multidisciplinary Research (IAJMR)
ISSN: 2454-1370

WEB PAGE DATA EXTRACTION AND CATEGORIZATION WEB PAGES USING RULE BASED CLASSIFICATION ALGORITHM

Nirav Prajapati[1], Ketan Modi[2] and Uttam Chauhan[3]

[1 &2]Merchant Engineering College, Basna, Mehsana, Gujarat, India.

[3]Vishwakarma Gov. Engg. College, Ahmedabad, Gujarat, India.

Abstract

A day, World Wide Web is the comprehensive information pool, but the data on internet is in unstructured or semi structured form and the useful data must be needed to be captured, stored and make in useful meaning. This poses a great challenge how to extract useful information from the web, which is in the form of unstructured or semi-structured and to identify the information is related to topic. Web page categorization is the process of categorizing web page into predefined categories based on their content. Here we are focusing on the extracting the unstructured web page data from URLs and other links, which are associated with URLs and stored them into database. This stored content remove stop words, repeated word and performing stemming on it and generate new dataset of those stored content. This dataset is matched with the pre-decided domain data dictionary and classifying according to the algorithm and matching rules for deciding it relevant category.

Key words: Web Data Extraction, Web Page Categorization, Web Page Classification and Web Analysis.

1.Introduction

Today, internet is a growing database people's lives are greatly influenced by information technology due to the excessive exposure of the Internet. The information available on internet is in three types of web pages, which are unstructured, semi-structured and structured web pages and there is a need for extracting useful information from it. By Merrill Lynch estimated that more than 85 percent of all business information existed as unstructured data. The Unstructured data such as multimedia files, presentations, documents, news, chats report and emails etc. are difficult to capture and store in the common database storage. Through effective unstructured data management, revenue, profitability, and opportunity can go up, while risks and costs may go down.

A. Data Extraction

The first step is to obtain the web page for data extraction. Data extraction is a process of retrieving or capturing the data from one medium to another medium. The medium can be documents, databases, repository, stack, or anything that consists of information. In the process of data extraction a basic operation of web crawler are perform in which it starts with the input typed web pages urls parsing and indexing. Create a copy of the visited web pages through the internet.

In order to extract data from a webpage two tasks need to be considered one is Input that can be unstructured page, semi-structured and

*Corresponding author: **Nirav Prajapati**
Tel.: +91 8128424713[1], +91 9723706169 [2], +91 9824897671[3]
E-mail: prajapatinrv13@gmail.com[1], modiketanit@gmail.com[2], ug_chauhan@gtu.edu.in[3]
Received: 15.03.2015; *Revised*: 07.04.2015; *Accepted*: 16.04.2015.

structured page while extraction target means data repository.

B. Data Cleaning

Data cleaning is also called data cleansing or scrubbing which are deal with detecting and removing errors and inconsistencies of data in order to improve the quality of data. During searching process some words which do not have important significant information or not any usefulness' such words (e.g., is, a, an, the, etc.) are stop words need to be removed from it. Finally, the text data is cleaned by removing unnecessary words or noisy data i.e. text data is filtered and subject related words are collected.

C. Page Classification

Web page classification, also known as web page categorization it is the process of assigning a web page to one or more predefined domain category. The goal of classification is to build a set of models that can correctly predict the class of different objects. Web page classification can also help to improve the quality of web search.

Web page classification can be divided into specific problems: subject classification, functional classification, sentiment classification, and other types of classification. Subject classification is concerned about the subject or topic of a web page. For example judging whether a page is about arts, business or sports subject. Functional classification cares about the role that the web page plays. For example, deciding a page to be a personal homepage, course page or admission page. Sentiment classification focuses on the opinion on particular topic that is presented in a web page. i.e. the author's attitude about some particular topic. Based on the number of classes in the problem, classification can be divided into binary classification and multi-class classification, where binary classification categorizes instances into exactly one of two class's multi-class classification deals with more than two classes. Based on the organization of categories, web page classification can also be divided into flat classification and hierarchical classification. In flat classification, categories are considered parallel, while in hierarchical classification, the categories are organized in a hierarchical tree-like structure.

Web pages can be classified by two methods: syntactic and semantic. This proposed work emphasizes syntactic classification, which uses a set of words or patterns in a web page to classify it.

2. Classification Methodology

Classifier systems takes an input each of them belonging to one of a small number of classes and it described by its values for a fixed set of attributes, and output of a classifier can accurately predict the class to which a new case belongs. Web Page Classification can done based on various parameters such as text, image, structure of documents etc. and by different feature selection methods like term occurrence number, term frequency, text content, link and content analysis and document frequency etc.

Many algorithms have been developed to deal with automatic text classification. The most common techniques used for this purpose include k-nearest neighbor classifier (KNN), Decision tree, Naive Bayes Classifier, Apriori Algorithm, SVM, FP Growth Algorithm and Rule based Algorithm.

A. K-Nearest Neighbor classifier (KNN)

K-Nearest Neighbor (KNN) classification is a non-parametric and very simple, yet powerful classification method. The key idea behind KNN classification is that similar observations belong to similar classes. It stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN algorithm usually use the Euclidean or the Manhattan distance.

The advantages and disadvantages issues of KNN are as follows

Advantages

- High accuracy.
- No assumption data.
- Easy to understand and implement classification technique.

- Perform well on application in which a simple and can have many class labels.

Disadvantages:

- Computationally expansive.
- Large storage requirements.
- Sensitive to the local structure of the data.
- Nearest neighbor classifiers assign equal weight to each attribute. This may cause confusion when many irrelevant attributes in the data and results into poor accuracy.

B. A Decision Tree

Decision tree builds classification or regression models in the form of a tree structure. It breaks a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is in form of tree with decision nodes and leaf nodes.

The advantages and disadvantages of Decision Tree Induction are as follows.

Advantages

- Decision Trees are very simple and fast.
- Working with continuous attributes.
- Performs well with large datasets.
- It does not require any domain knowledge or parameter setting.

Disadvantages

- It has long training time lack of available memory, when dealing with large databases.
- Small variations in the data can imply that very different looking trees are generated

Examples: The popular decision tree algorithms are ID3, C4.5, CART.

C. Naive Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on Bayes theorem with independence assumptions between predictors. It provides a flexible way for dealing with any number of attributes or classes, based on probability theory of Bayes rule.

The advantages and disadvantages of Naïve Bayesian are as follows.

Advantages

- It does not require large amounts of data before learning can begin.
- Easy interpretation of knowledge representation.
- Naive Bayes classifiers are computationally fast when making decisions.

Disadvantages

- Less accurate compare to other classifier.
- May not be best classifier in any particular application, but it does well and robust.

3. Discussion

In this research we are discussing about the categorizing the web pages from URL in to particular class. For categorization use a datasets which contain the word list of URL web page and domain class dictionary in which the web page is being categorized. Generate a domain class data dictionary for comparing with the dataset and categorized web pages. The more amount of words would there be in the dictionary, the more accuracy will be generated in data processing. Data extraction from URL uses a crawler it extract all the web page text data from URL and store them into database and transform into the text file. From this above text file perform the data cleaning operation for creating uniquely datasets.

For data cleaning operation we need to perform the removing the stop word i.e. the, is, an etc. from text file and generate a new important meaningful wordlist known as "Dataset". The generated dataset also contain some repeated no, of words so it also increases the complexity of process so it is necessary to removed from datasets. After these perform the operation of removing repeated word from datasets and generate new word. This datasets contain some of the stemmer word, we need to perform the stemming operation on it through stemming operation similar meaning of word are removed

and it creates only one word from those similar word through this process accuracy and performance speed of process is increased. At last a new uniquely important meaning full datasets from text file generates.

At finally compare this datasets file with the domain class dictionary and according to the matching rules of rule based algorithm and pre decided threashold of matching list the web page classify into its relevant category.

Following are the steps for extracting web data and classifying in categories.

Step - 1 take the some URLs as an input.

Step - 2 uniquely defines all the links in new lists.

Step - 3 extract all the text data from URLs and store in the database.

Step - 4 remove all the stop word and repeated words from the stored contents and generate a new dataset.

Step - 5 perform stemming operation on datasets and generate a new important significant word set of that datasets.

Step - 6 compare this uniquely generated datasets with pre-defined domain data dictionary and create new datasets of compare data.

Step - 7 calculate the frequency of matched data with the dictionary and according to the rule based algorithm and matche word threashold value categorize web page in relevant domain class.

4. Conclusion

Here we can conclude that the generated search which are associated with URLs are firstly stored into database to give the appropriate search. This stored content remove stop words, repeated word and performs stemming on it and generate a new dataset entry of those stored content. This dataset is then matched with the pre-decided domain data dictionary and classify according to the algorithm and matching rules for deciding it relevant category and give the user a desired result. By doing this the web analysis possibility becomes more accurate and the results are more specifically generated.

5. References

1) Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh ·Qiang Yang ·iroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg. 2007. "Top 10 algorithms in data mining" At © Springer-Verlag London Limited.
2) Xiaoguang Q. I and Brian D. Davisonacm. 2009. "Web page classification: Features and algorithms". Computing Surveys, Vol. 41, No. 2, Article 12, Publication date: February.
3) Sini Shibu, Aishwarya Vishwakarma and Niket Bhargava. 2010. "A combination approach for Web Page Classification using Page Rank and Feature Selection Technique". International Journal of Computer Theory and Engineering, Vol.2, No.6, December.
4) Siti Z. Z. Abidin, Noorazida Mohd Idris and Azizul H. Husain. 2010. "Extraction and Classification of Unstructured Data in WebPages for Structured Multimedia Database via XML" ©2010 IEEE
5) Hetal Bhavsar and Amit Ganatra. 2012. "A Comparative Study of Training Algorithms for Supervised Machine Learning", IJSCE ISSN: 2231-2307.
6) Lambodar Jena and Narendra Kumar Kamila. 2013. "Data Extraction and Web page Categorization using Text Mining" International Journal of Application or Innovation in Engineering & Management (IJAIEM) ,Volume 2, Issue 6.
7) Abdelhakim Herrouz, Chabane Khentout and Mahieddine Djoudi. 2013. "Overview of Web Content Mining Tools". The International Journal Of Engineering And Science (IJES), Volume-2, Issue-6.
8) Mohd Fauzi Bin Othman, Thomas Moh Shan Yau, J. Han and M. Kamber. 2011. "Comparison of Different Classification

Techniques Using WEKA for Breast Cancer". Data Mining Concepts and Techniques, Elevier.

9) Yi Cheng, Jianye Ge, Jun Liang and Sheng Yu "Comparison of Web Page Classification Algorithms".
10) Keyur J. Patel. 2013. "Web Page Classification Using Data Mining". International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 7.
11) Sonal Vaghela, M. B. Chaudhary and Devendra Chauhan. 2014. "Web page classification using term frequency". International Journal For Technological Research In Engineering Volume 1, Issue 9.