Web Page Feature Extraction And Classification Using Neural Network

Shantanu Kakade[1], A. J. Dasare[2]
[2] Professor
[1, 2] Vishwakarma Institute of Technology

Abstract- *Majority of information systems today such as search engines, recommendation systems and other artificial intelligence systems gather information from the internet. Most of the information on the internet is in the form of web pages. To effectively and efficiently process this information, there needs to be a system which will categorize these web pages based on their content. The features considered for classification are the structure of the home page, the position of text, images and other data on the page, the amount of internal and external links, frequency of certain buzzwords, etc. The web page is represented in the form of a DOM (Document Object Model) tree. Each node of the DOM tree represents a data item on the web page.*

Current classification systems need to be calibrated to a particular web page template before they can analyze its content. These systems can classify only those web pages which fit the template on which the system was trained on. In our system, we implement neural network which can analyze the content of a web page regardless of its structure. This is done by training the neural network on different types of web pages, with varying content and structure. When a new web page is given to the network, it finds the position of nodes in the DOM tree and calculates the probability that the given web page belongs to one of the predefined classes.

Keywords- Web page feature extraction, Neural Network, DOM tree, classification

I. INTRODUCTION

Web page classification plays an important role in many data management and retrieval jobs. On the internet, categorization of page content is crucial to crawling, to the supported development of web directories, to content-specific URL analysis, to context based advertising, and to analysis of the structure of the page. Web page classification also assists in improving the quality of web search.

Feature extraction of a web page refers to the separation and identification of various elements of the page such as text, images, hyperlinks, animation, etc. The features of a web page represent its category. To determine the category of a web page, its features are analyzed according to their content, position and frequency.

The classification of web pages is based on the fact that pages of a similar type, i.e. those that belong to the same category have similar features. Therefore, the content, position and frequency of the page's features is going to be similar. We use this information to put the web page into one of the predefined categories.

A. Motivation

The internet is the biggest and fastest growing source of data in today's world. Majority of information on the internet is in the form of web pages. Many information systems need to gather data from the web pages. Therefore categorizing of web pages is getting increasingly necessary.
Due to the large and ever-increasing amount of data on the internet, manually classification of web pages is not practically possible. Hence automatic categorization of web pages is necessary to assist the end users as well as the web servers and search engines to find the desired web sites. Such a system can help other information system to automatically sort through large amounts of data and select the most relevant content from the web pages. The aim is to create a system that can automatically extract the features from a web page and categorize it.

B .Objective

- To collect the dataset of web pages with sufficiently varying content. Since we need the system to be independent of the structure of the web page, we will need to train it by using a number of different styles of web pages.
- To create and train a neural network which can identify the class of a web page based on factors like the position of data items in the DOM tree, the text content, number of internal and external links, etc.
- Test the neural network on new data and measure the accuracy of classification. Make necessary changes to the network architecture if the overall accuracy is less than expected.

- Use this system to assist end users as well as other information systems like search engines, recommendation systems , web servers to help them categorize online data

II. LITERATURE REVIEW

Sr No	Title	Author	Conference	Methodology
1	Deep Neural Networks for Web Page Information Extraction	Tomas Gogar, et al.	IEEE	Page classification using Convolutional Neural Network
2	Web Page Categorization Using Artificial Neural Networks	Sikder M. Kamruzzaman	International Conference of Electrical Engineering	Page classification based on text content, external links and animations
3	The Automatic Classification of Web Pages Based on Neural Network	Yizhong Zhang, et al.	IEEE	Using Self-organizing feature map for page classification
4	Automatic Classification of Web Pages using Back Propagation	Poonam Nagale, Arti Waghmare	International journal of Computer applications	Use of backpropogation in web page classification
5	Web Content Extraction Through Machine Learning	Ziyan Zhou, Muntasir Mashuq	ACM SIGMOD Record,	Classification of unstructured data
6	Fast Webpage Classification Using URL Features	Min-Yen Kan Hoang Oanh Nguyen Thi	ACM 1-59593-140-6/05/0010	Using URL features to determine category

III. METHODOLOGY

A. Using CNN for classification

- In this system, we use neural network to learn the features of web pages and classify them accordingly. The web pages are represented in the form Document Object Model (DOM) tree. Each node in the DOM tree is an element on the web page such as a text paragraph, image, hyperlink, multimedia content, etc.
- To effectively categorize the web page, we need to first define the categories of web pages. For this system, we define some of the most common categories of web pages such as Ecommerce, Education, News, Entertainment and Science. The factor that distinguishes one category from another is the content and positioning of nodes in the DOM tree. For example, Ecommerce web sites have a roughly similar template where the product image is at the top of the page, followed by a description of the product, then a price label as well as multiple hyperlinks to other pages of the same site throughout the page. News websites typically have a heading and a large body of text interspersed with images. We train the neural network to detect these patterns in the DOM tree so as to recognize different categories.
- Existing system need to be calibrated to the structure of a web page before they can analyze it. Since we need our system to be independent of the structure, we train it to detect the relative position of nodes in the DOM tree in the web page. This allows the neural network to identify the class of a web page based on the relation between its features, instead of relying on the web page structure.

B. *Categories*:

- We have selected common website categories such as ecommerce, education, news, entertainment and science. The neural network is trained on dataset of web pages from these categories.
- For each of these classes, sub-classes of features are defines. For example, an ecommerce webpage has product image, price label, ‘Add to cart’ button, etc. The neural network learns to detect these sub-classes and the features of web pages are assigned to them.

C. Representation of Web page using DOM tree

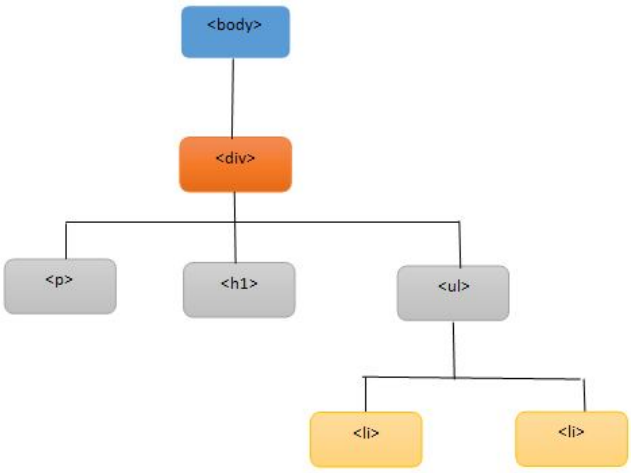

Figure 1 Example DOM tree

- The Document Object Model (DOM) is a programming API for HTML and XML documents. It defines the logical structure of documents and the way a document is accessed and manipulated. It is a cross-platform and language-independent application programming interface that parses an HTML, XHTML, or XML document as a tree structure wherein each node is an object representing a part of the document. The objects can be manipulated programmatically and any visible changes occurring as a result may then be reflected in the display of the document.
- To render a document such as an HTML page, most web browsers use an internal model similar to the DOM. The nodes of every document are organized in a tree structure, called the DOM tree, with the topmost node named as "Document object". When an HTML page is rendered in browsers, the browser

downloads the HTML into local memory and automatically parses it to display the page on screen.

D. Architecture

- Input to the network

1. Since the neural network only takes numerical values as input, the DOM tree cannot be given to the network as it is. For each node in the DOM tree, its position in the web page is calculated as x and y co-ordinates. A numerical value is assigned to the type of node. This value, the x and y co-ordinates and other information, such as whether the node contains a hyperlink, is given as input to the network.
2. To process the text content, the text nodes are converted to vectors and given input to the network.
3. The number of external and internal hyperlinks in the document are given as input to the network.

- Network architecture

1. We use a feed-forward neural network with backpropogation algorithm. The number of neurons in the input layer corresponds to the number of numerical inputs to the network
2. The number of neurons in the hidden layer will be determined based on trial and error approach and it will be updated if the overall accuracy of the system is less than expected
3. The output of the network will represent which category the web page belongs to. Hence the output layer will have 5 neurons, to signify 5 predefined categories.

- Algorithm:

1. Collect sample dataset of web pages of different categories and structure.

2. Feature Extraction:

   a. Obtain the DOM tree of the web page
   b. Assign a numerical value to the type of nodes in the DOM tree
   c. Calculate the position of nodes (in pixels) in the web page
   d. Find out if the node contains a hyperlink or not
   e. Calculate the total number of internal and external links in the web page
   f. Generate text vectors for the textual content
   g. Define sub-classes of features for all categories
   h. Assign a numerical value to the class of web page

3. Training

   a. Create neural network with required architecture
   b. Provide the sample dataset of web page with their category.
   c. Perform pattern matching between provided dataset and its category
   d. Train the network until the % error is below a certain threshold

4. Evaluation

   a. Provide a new set of web pages whose category is to be determined
   b. Analyse the features of given web page
   c. If a node is away from the centre of page and has a hyperlink to another website, mark it as an advertisement and ignore that node.
   d. Classify the page into one of the predefined categories

IV. RESULTS

- Dataset

Category	No. of pages
Ecommerce	5
Entertainment	4
News	4
Science	3

- Training Accuracy

No. of Epochs	Avg Accuracy
1000	80.40%
1500	81.02%
2000	78.30%
3000	80.32%

V. CONCLUSION

This system provides a new way for categorizing the web pages on previously unseen data. Since the system does not need any site specific initialization, it can classify a web

page of any structure or template. In this way, the system is more intelligent than previous systems.

The need for manual recalibration is eliminated, improving the efficiency of the system,

Since the system is trained to detect irrelevant data, such as ads, it can ignore such data, thus improving accuracy and reducing processing time since the unnecessary does not get analysed for classification. This eliminated the need for manual pre-processing of data and making the system more autonomous. Also, by using multiple features of the web page for analysis, the system is more accurate in its classification than the systems which use consider only one type of extracted feature for classification. In future, the system can be trained to identify more number of categories, as well as more types of features such as images, videos, animations, etc. The system can also be combined with image processing techniques for processing the visual data in the web pages.

REFERENCES

[1] Tomas Gogar(B), Ondrej Hubacek, and Jan Sedivy, "Deep Neural Networks for Web Page Information Extraction", Springer, 2016
[2] Yizhong Zhang, Mingsheng Zhao, "The Automatic Classification of Web Pages Based on Neural Network", 2008
[3] Sikder M. Kamruzzaman, "Web Page Categorization Using Artificial Neural
Networks", International Conference on Electrical Engineering, 2010
[4] Poonam Nagale, Arti Waghmare, "Automatic Classification of Web Pages using Back Propagation", International Journal of Computer Applications, 2013
[5] Baudiˇs, P., ˇSediv´y, J.: Sentence pair scoring: towards unified framework for text comprehension. ArXiv preprints, March 2016
[6] Fan, S., Wang, X., Dong, Y.: Web data extraction based on visual information and
partial tree alignment. In: 2014 11th Web Information System and Application
Conference (WISA), September 2014, pp. 18–23 (2014)
[7] Ferrara, E., Meo, P.D., Fiumara, G., Baumgartner, R.: Web data extraction, applications and techniques: a survey. Knowl. Based Syst. **70**, 301–323 (2014)