Web Content Extraction Using Machine Learning

Shwetangi Gurav[1], Jahir Gilani[2],Vinit Gore[3], Prof. Shilpa Jadhao[4]

[1,2,3] Bachelor of Engineering, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India
[4]Assistant Professor, Dept. of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India

---***---

Abstract: *The World Wide Web has seen tremendous growth in recent years. With the large amount of information on the Internet, web pages have been the potential source of information retrieval and data mining technology such as commercial search engines, web mining applications. Internet web pages contain several items that cannot be classified as the informative content, e.g., search and filtering panel, navigation links, advertisements, and so on called as noisy parts. Most clients and end-users search for the informative content, and largely do not want the non-informative content. A tool that assists an end-user or application to search and process information automatically ,must separate the "primary or informative content sections" from the other content sections. The content extraction problem has been a subject of study ever since the expansion of the World Wide Web. Its goal is to separate the main content of a web page, such as the text of a news story, from the noisy content, such as advertisements and navigation links. Most content extraction approaches operate at a block level; that is, the web page is segmented into blocks and then each of these blocks is determined to be part of the main content or the noisy content of the webpage. The extracted main content is summarized into tabular format.*

Keywords: **Webpages, Crawler, Clustering, K-Means, SVM.**

1. INTRODUCTION

The web pages (also referred to web documents) that constitute the World Wide Web are sources of very diverse categories of information. These include news, reference materials, forum discussions, and commercial product descriptions, just to name a few. Each category of information can in turn have various media formats, such as textual, graphical, or video. This vast amount of information is used by ordinary web users throughout the world, as well as by automated crawlers that traverse the Web for various purposes, such as web mining or web indexing.

In most cases, however, a single webpage consists of distinct "parts," which will be referred to in this thesis as the contents of the webpage. Only one type of content, which will be referred to as the main content of the webpage, is what makes the webpage a useful source of information. Other contents include advertisements, navigation buttons, page settings, and legal notices; these contents will be collectively referred to as the noisy content of the webpage. The process of identifying the main content of a web page is called main content extraction, or more briefly content extraction.

The goal of this thesis is to induce new rules for content extraction using supervised machine learning algorithms based on a sample of webpages with manually labeled contents; that is, the contents of these webpages have been identified as main or noisy by a human annotator. In addition, the content extraction performance under these rules should be evaluated.

2.LITERATURE REVIEW

SR. NO.	TOPIC	YEAR	AUTHOR	METHOD	SHORTCOMINGS
1	Noise Removing from Web Pages Using Neural Network	2010	Thanda Htwe, Khin Haymar	Content extraction using Tag path clustering	Css properties are not considered during clustering.
2	DOM based content extraction via text density	2011	Fei Sun, Dandan Song, Leijian Liao	Calculating Text Density after content extraction	Irrelevant data was not removed during content extraction.
3	Noise reduction and content retrieval from webpages	2013	Surabhi Lingwal	Web content extraction using data mining	Less precision compared to machine learning.
4	Web Usage Mining: A Review on Process Methods and Techniques	2013	Chintan Varnagar, Nirali Madhak	Data mining approach that uses log server files	Log usage requires memory and also creates redundancy.

3.SYSTEM ANALYSIS

3.1.PROBLEM STATEMENT

Web content extraction is a key technology for enabling an array of applications aimed at understanding the web. While web extraction has been studied extensively, they often focus on extracting structured data that appear multiple times on a single webpage, like product catalogs. This project aims to extract less structured web content, like news articles, that appear only once in noisy webpages. Our approach classical text blocks using a mixture of visual and language independent features. In addition, a pipeline is devised to automatically label data points through clustering where each cluster is scored based on its relevance to the webpage description extracted from the meta tags, and data points in the best cluster are selected as positive training examples.

3.2.PROPOSED SYSTEM

Proposed approach concentrates on web pages where the underlying information is unstructured text. The technique used for information extraction is applied on entire webpage and they actually seek information only from primary content blocks of the web pages. The user species is required information to the system. Web crawlers download web pages by starting from one or more seed URLs, downloading each of the associated pages, extracting the hyperlink URLs contained therein, and recursively downloading those pages.

3.3.WORKING

The system contains a main website in which the URL of the webpage whose data needs to be crawled is entered and the flow of the system goes as follows:

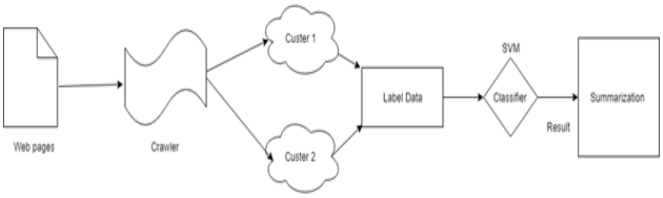

Figure 1: System Flow

1) Get URL of a website as input

2) Crawler will crawl the site and extract text data

3) Apply clustering on extracted data to divide data in clusters like text, links, etc.

4) Labeling the data by finding out related data to that page as 1 and unrelated data as 0.

5) Apply SVM to classify data as content and non-content. This process will remove noise, adds, duplicated data.

6) Final output will be the summarized text from website.

4.CONCLUSION AND FUTURE SCOPE

Our pipeline collects data, labels examples, trains support vector classifier, and evaluates learned model in an automated manner. Our learning algorithm can achieve perfect labeling when trained on a single website.
We further aspires to include these changes:

1. The summarization is performed in tabular format. Instead of displaying links of images, the description of the images can be displayed.

2. Furthermore these links to the content of the data can be made active for the reference of human users.

3. The system can be scaled to work on a website as a whole, and even further for multiple websites. However, the memory required and the cost associated for the same will be very high. For commercial implementation, this option is worth trying.

4. Again from human user's experience concern, the clusters of text, images and links can be themselves separately shown. A continuous access to all images in a sequential manner can also be shown.

ACKNOWLEDGEMENT

This research was supported by our project guide Ms. Shilpa Jadhao, We thank her for providing us an opportunity to do the project research on Web Content Extraction Using Machine Learning at Rajiv Gandhi Institute of Technology (R.G.I.T.), Andheri, Mumbai and by giving us the support and guidance.

REFERENCES

[1] Ziyan Zhou and Muntasir Mashuq. Web content extraction through machine learning, 2014. 1.1, 3.8.1

[2] Johnny Stenback, Philippe Le Hégaret, and Arnaud Le Hors. Document object model (dom) level 2 html specification. W3C Recommendation, 2003. 2.2

[3] Pang-Ning Tan et al. Introduction to data mining. Pearson Education India, 2006. 4.1.2

[4] Terry Therneau, Beth Atkinson, Brian Ripley, and Maintainer

[5] Brian Ripley. Package 'rpart', 2015. 4.1.1

[6] Jae-Woo LEE "A Model for Information Retrieval Agent

[7] System Based on Keywords JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN COMPUTER ENGINEERING ISSN: 0975 – 6760| NOV 12 TO OCT 13 | VOLUME – 02, ISSUE –

[8] 02 Page 297 Distribution" International Conference on Multimedia and Ubiquitous Engineering(MUE'07), IEEE 2007 0-7695-2777-9/07

[9] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a competition for cleaning web pages. In LREC, 2008.