JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 1, 2020

Review on Web Page Data Extraction Technique

Hoor Fatima^1, Gaurav Raj^2, Sudeshna Chakraborty^3

[1,2 ,3]Dept. of Computer Science and Engineering, Sharda University, Greater Noida, U.P.

Email Id- [1]hoor.fatima@sharda.ac.in, [2]gaurav.raj@sharda.ac.in, [3]sudeshna.chakraborty@sharda.ac.in

Received: 02 November 2019 Revised and Accepted: 02 January 2020

ABSTRACT: By the fast growth of Communications knowledge and communication growth, in only Internet filled despite junk, also websites used as straightforward for navigation pane. Mainstream web page's text data typically has granular permissions and could be separated interested in a short script sheet and a multiple-script sheet. To retrieve the webpage script data, the location of textual details can be identified correctly using the multiple text functionality and webpage design rules. Web page data extraction means to collect data from a website by submitting a link to the requested page, then combing different objects through the Markup and sorting the data. This article suggested a technique for obtaining site page script data founded on multiple-function synthesis, according to the above features. Simulations based on a huge amount of information demonstrates that the method is standardized and highly accurate for extracting text information from solitary text, multiple-text webpages and also actual appropriate for net applications.

KEYWORDS: Text data extraction, Multi-factor fusion, Multiple-text body webpage, Solitary script body webpage.

I. INTRODUCTION

As Internet technology evolves quickly and information expands, the Internet these days is not only inundated with spam, websites are not as easy as the earlier pages. The sheet includes several elements including screen types, texts, and lots of commercials. In what way to discovery some interesting data on various junk and identify the position of subject knowledge on the website correctly and absolutely that in current research partakes now converted a blistering topic. [1], [2] There have been lots of work in this field and many thoughtful techniques in the area of internet-sheet text knowledge processing. The data source from either the Web page derived out of the same domain or net pages is focused on architecture of the DOM tree and other enhancement techniques. The derived the data collection on a site is not only a single official site focused on visual, mathematical theory-driven numerical data, etc. Many writers merge this two forms of expertise, like Road Runner. [3]

Determining the location of text details using the textual functionality of the website primarily requires statistical theory: It used the body three characteristics as information (pointing, non-hypertext and hyperlink) translated to statistical data values to decide the location of body knowledge, used and imprinted Song's process. [4]

Throughout reality, today's approaches based on statistical analysis are restricted. The consolidation of website types limits the precision of selection and does not have a high flexibility. This paper intends to develop a better framework for collecting and structuring text data for specific sort of net pages. [5]

The program is ideal aimed at web pages and websites in various styles.

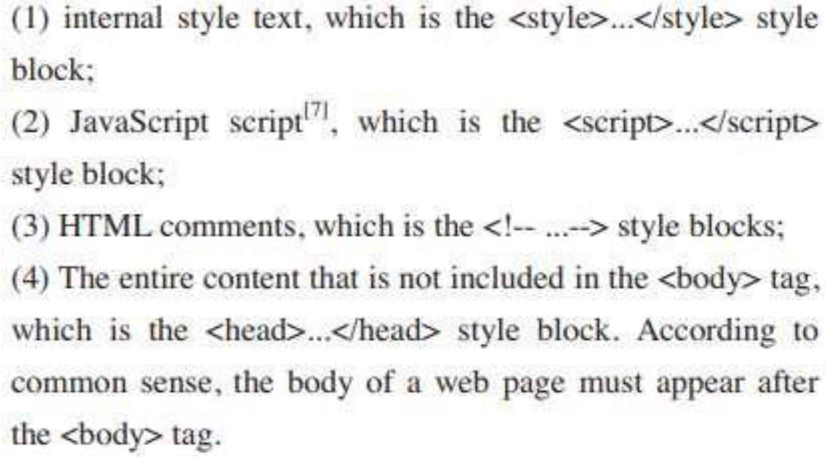
The great precision and flexibility of result obtained is a hard scenario in development of the text abstraction algorithm for the website. We research the process for the extraction of website text data with high accuracy and flexibility centered upon the Identifiers the well-known web browsers. [6]

A multi-channel webpage knowledge extraction approach (WFFTE) is suggested in direction to satisfy complexity of the net page sort and simplicity of this process alone. [7]

II. REVIEW

To increase the transfer speed, it must be eliminated earlier the label tree is set up to advance storage performance any noticeable noise in the origin document like HTML comments and scripts.

Standard words are used to search and evaluate HTML documents for noise data. Further information regarding the caption of the text, the document account, the text script used throughout the text, its style meaning, the text term is included in the < HEAD > flag. The data that the client will access is located in the section < BODY > in the section < BODY >, the body information is kept. The report also discusses only the < BODY > contents. A large number of HTML reports have been systematically evaluated so some apparent noise analysis in the website are mainly:



(1) internal style text, which is the <style>...</style> style block;

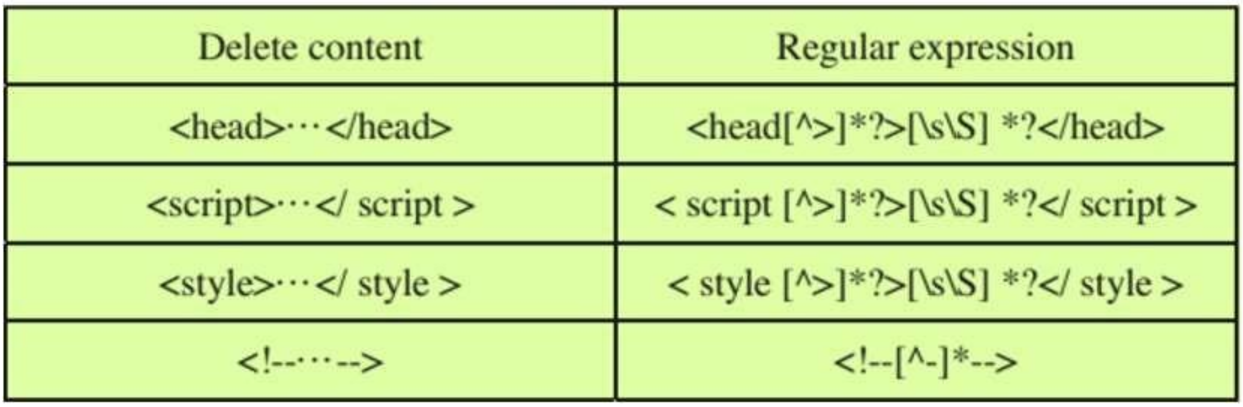
(2) JavaScript script[7], which is the <script>...</script> style block;

(3) HTML comments, which is the <!-- ...--> style blocks;

(4) The entire content that is not included in the <body> tag, which is the <head>...</head> style block. According to common sense, the body of a web page must appear after the <body> tag.

Table 1 displays the storage system. Experiments show that the papers were reduced to about 30 percent after denotation using regular expressions, rendering subsequent work considerably easier.

Table 1: Details and Steps for Deleting Redundant Tags



Delete content	Regular expression
<head>⋯</head>	<head[^>]*?>[\s\S] *?</head>
<script>⋯</ script >	< script [^>]*?>[\s\S] *?</ script >
<style>⋯</ style >	< style [^>]*?>[\s\S] *?</ style >
<!--⋯-->	<!--[^-]*-->

The key features of the multiset web pages are to spread the material via several container tags and the presentation style of these container tags, which may also include several dot lines, is likely to be the same depending on the graphical patterns of web design. The document also includes several name details, next to the title mark. To sum up, evaluating the functionality of various web sites, the body of web content has a quantity of structures like text quantity, corporal punctuation, code hyper-link script, non-hyperlink script, descriptive language figure of name, the length from the title and body of knowledge type and position. [8]

The suggested WFFTE is to transform the HTML file in a DOM diagram and to measure the script assistance of every vessel tag by comparing the text functionality of these web pages. The optimization steps are defined together with some computation in the computation phase and after calculation. [9]

This paper uses Jsoup .jar third-party jar kit to create the web page DOM tree. The jar kit has the task of first correcting the standard web page tags, and then parsing the xml file to cross all tabs with a markup tag as the root node. Figure 1 shows the DOM list. [10]

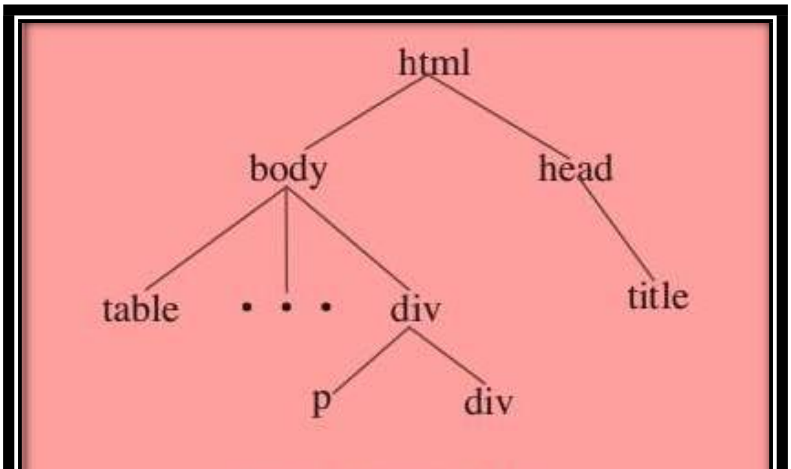


Fig.1. Web page DOM tree

The special characteristics of every component tag location direction are obtained by crossing the DOM tree by uniquely marking each container mark, e.g. by identifying the Figure 1 DOM tree and having the outcomes as shown in Figure 2.

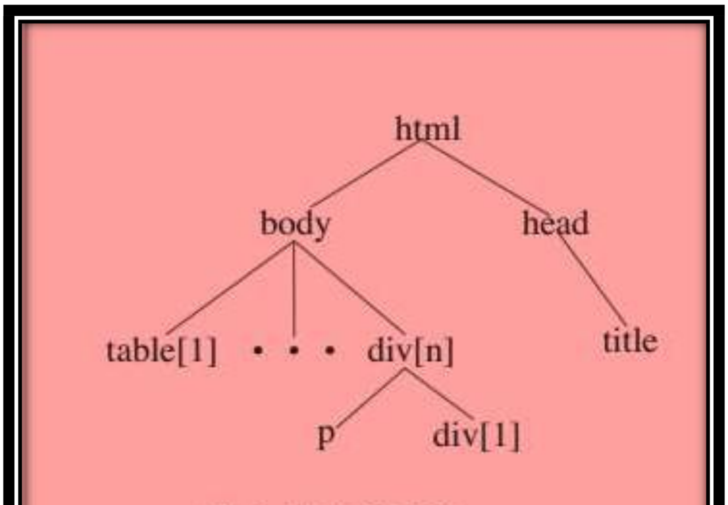
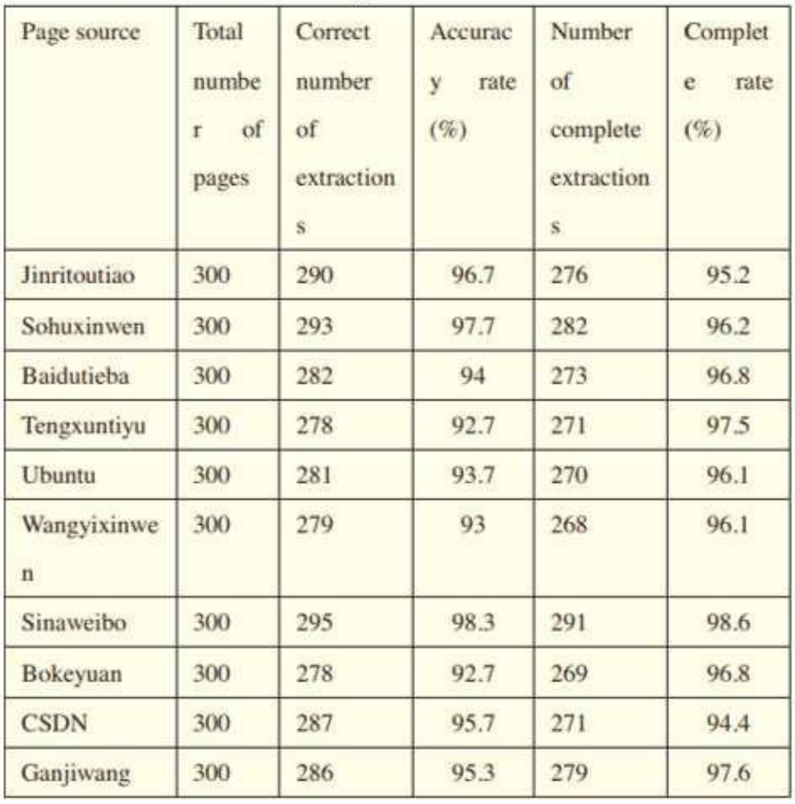


Fig.2. labeled DOM tree

III. RESULT

Table 2: Experimental Result



Page source	Total number of pages	Correct number of extractions	Accuracy rate (%)	Number of complete extractions	Complete rate (%)
Jinritoutiao	300	290	96.7	276	95.2
Sohuxinwen	300	293	97.7	282	96.2
Baidutieba	300	282	94	273	96.8
Tengxuntiyu	300	278	92.7	271	97.5
Ubuntu	300	281	93.7	270	96.1
Wangyixinwen	300	279	93	268	96.1
Sinaweibo	300	295	98.3	291	98.6
Bokeyuan	300	278	92.7	269	96.8
CSDN	300	287	95.7	271	94.4
Ganjiwang	300	286	95.3	279	97.6

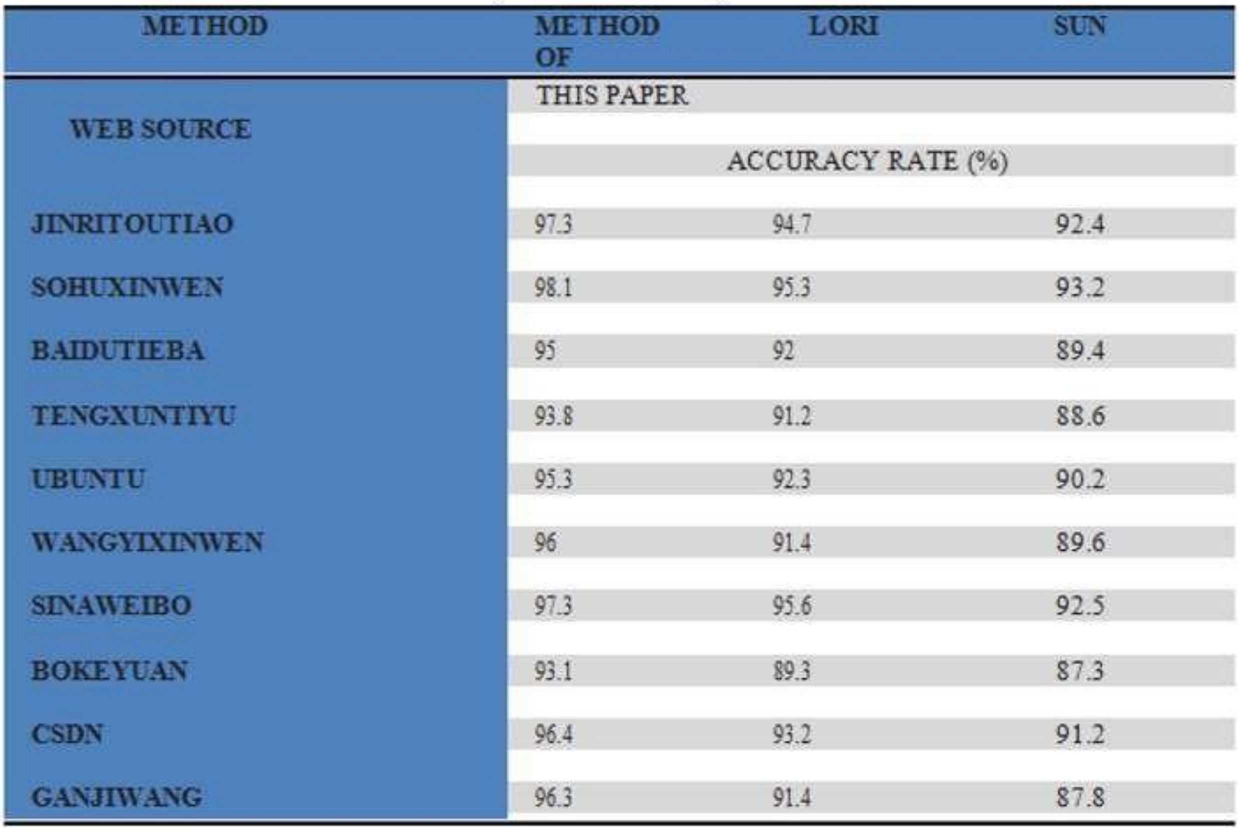
The study featured 10 websites: Ganjiwang, Jinritoutiao, Sohuxinwen, Baidutieba, Tengxuntiyu and Ubuntu. 200 web pages have been picked by chance from these web pages, and Table 2 displays mathematical models. To determine the mathematical models, utilize the exactness P, clarity's R of derived writing data and its formula of measurement is as follows:

$$P = \frac{C2}{C1} \times 100\%$$

$$R = \frac{C3}{C2} \times 100\%$$

Where the highest limit of web sites in the analysis is C1, C2 is correctly collected the lot of web sites containing text, and C3 the number of online pages containing full textual data. The accuracy rating is dependent on the overall variety of online pages and the whole rating on the variety of online pages with correctly interpreted text information. [11]

Table 3: Experimental Comparison Results



METHOD	METHOD OF THIS PAPER	LORI	SUN
WEB SOURCE	ACCURACY RATE (%)		
JINRITOUTIAO	97.3	94.7	92.4
SOHUXINWEN	98.1	95.3	93.2
BAIDUTIEBA	95	92	89.4
TENGXUNTIYU	93.8	91.2	88.6
UBUNTU	95.3	92.3	90.2
WANGYIXINWEN	96	91.4	89.6
SINAWEIBO	97.3	95.6	92.5
BOKEYUAN	93.1	89.3	87.3
CSDN	96.4	93.2	91.2
GANJIWANG	96.3	91.4	87.8

In this text, the technique for estimating the location details in the terminology is comparable by the LORI, SUN approaches [4], [5]. Table.3 above displays the investigational outcomes. As may be seen in Table 3, the precision with which websites with a single text and multi-textures can be retrieved is greater than traditional techniques.

Table 3 sows the greater precision of the method here. The data extraction tool, which is very useful for retrieval, is being applied to the website more and more perfectly usable variables. The product of removal of the website for a single text and the website for a multi textual type is very small.

IV. CONCLUSION

The approach based on large-scale design does not separate the experimental results from the actual application that can fulfill other scientific and functional objectives. The webpages of several texts have a marginally weaker extraction impact than single texts.

Such study is because the labels in the network layout are intricately interlocked, or they have different display formats and lengths. The texts on the webpages are various, and in fact the multi-text structure has its own unique nature, so in the future the algorithm can be further researched and strengthened.

V. REFERENCES

[1] Y. Zhai and B. Liu, “Web data extraction based on partial tree alignment,” 2005.

[2] Y. Zhai and B. Liu, “Extracting Web data using instance-based learning,” in World Wide Web, 2007.

[3] N. Jindal and B. Liu, “Analyzing and detecting review spam,” in Proceedings - IEEE International Conference on Data Mining, ICDM, 2007.

[4] Y. Zhai and B. Liu, “Structured data extraction from the web based on partial tree alignment,” IEEE Trans. Knowl. Data Eng., 2006.

[5] I. A. Ahmad Sabri, M. Man, W. A. W. Abu Bakar, and A. N. Mohd Rose, “Web Data Extraction Approach for Deep Web using WEIDJ,” in Procedia Computer Science, 2019.

[6] D. J. Hand, “Matrix Methods in Data Mining and Pattern Recognition by Lars Eldén,” Int. Stat. Rev., 2007.

[7] S. P. Algur and P. S. Hiremath, “Extraction of flat and nested data records from web pages,” Conf. Res. Pract. Inf. Technol. Ser., 2006.

[8] T. J. Bovend’Eerdt, M. Newman, K. Barker, H. Dawes, C. Minelli, and D. T. Wade, “The Effects of Stretching in Spasticity: A Systematic Review,” Archives of Physical Medicine and Rehabilitation. 2008.

[9] S. A. Gabarro, “Introduction to HTML,” in Web Application Design and Implementation, 2015.

[10] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, “DOM-based content extraction of HTML documents,” in Proceedings of the 12th International Conference on World Wide Web, WWW 2003, 2003.

[11] D. DiFranzo et al., “The Web is My Back-end: Creating Mashups with Linked Open Government Data,” in Linking Government Data, 2011.