Web Information Extraction: Tag Density and Keyword Approach

Shikha Shukla, Nitin, Sitendra Tamrakar
NRI Institute of Science and Technology

ABSTRACT

Web page consists of lots of noise in the form of advertisements, irrelevant information, copyrights information and menus. To extract the information from web we use the two concepts, text density and title of the page. Generally the main content of the page is denser than the other and noises has lesser text information. The title is the most important information on the page that tells us about what is this page for. So we simply extract all the information that is denser than particular threshold or at least contain one of the keywords that is made from the title of the page. By using this approach the more false negatives can be avoided. This approach gives very satisfactory results.

Keywords- Crawler, Web mining, information extraction

1. INTRODUCTION

The data on the web are increasing exponentially due to rapid use of social networking and e-commerce websites and users are relying more on the web for their daily activities such as online news, social networking, movies, shopping etc. As the more and more web users are increasing so does the noise like advertisements, bogus information etc. So it's become very important task to filter the noises and extract only those which are important.

Several works has been done in this field. Shin, Kwangcheol, and Geun Sik Jo [1] used style sheets to extract the information but the problem with this approach is it required the user interaction and user has to select some information on the page that he is interested in then the approach will search all the information that follow the same style sheets. Sun, Fei, Dandan Song, and Lejian Liao [2] used DOM tree based approach and extracting maximum density data as a main content. Asfia, Mohsen, Mir Mohsen Pedram, and Amir Masoud Rahmani [3] has used VCE(Visual clustering extractor) algorithm that uses DOM tree as input and produces smaller blocktree and some general parameters which are used to determine main block later. Downey, Doug, et al [4] extracted the information by adding the pattern learning algorithm which learns about the most common patterns in which instances of class appear. Yi, Lan, and Bing Liu [5] proposed cleaning technique is based on layouts and contents of the Web pages in a Web site. In their proposed method they first find a suitable data structure to capture and represent common layouts or presentation styles in a set of pages of the Web site and used compressed structure tree (CST) for this purpose. The compressed structure tree has some entropy measure assigned to the node which is used for the noise removal from the web page. Our approach uses the text density and combination of keywords to efficiently extract the information. This paper is divided is divided into three parts first is introduction which is almost discussed the next one is methodology which covers all the methods and concept to extract the information. And the last part is result evaluation which will show the performance of our approach.

2. METHODOLOGY

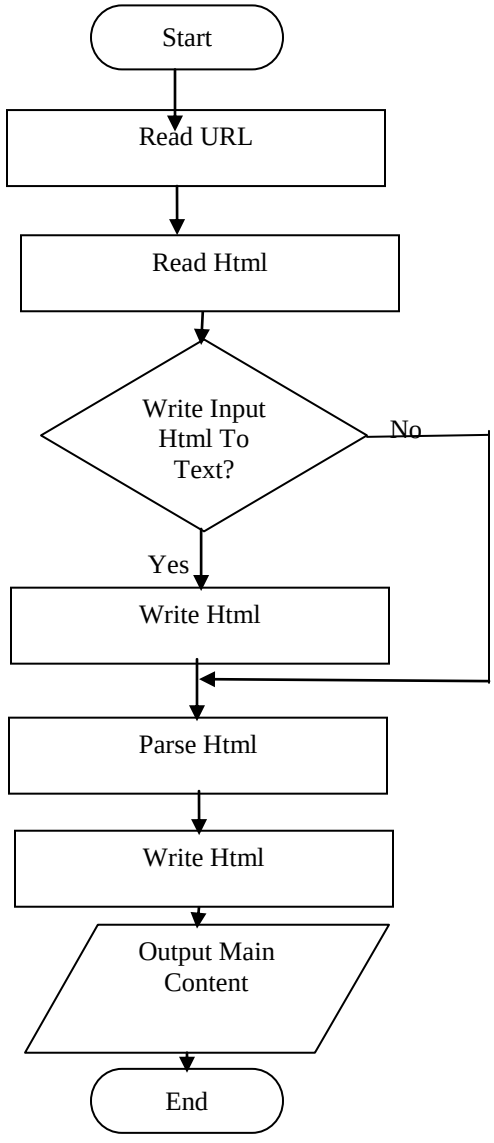


Figure 1 Flow Chart of Proposed Method

The flow graph of our proposed method is given in the figure 1 which shows that first desired URL is read and html page is loaded into memory which can be either written on the disk or directly passed to the main function Parse Html which takes the html page as input and produces the output HTML which is then written on the disk.

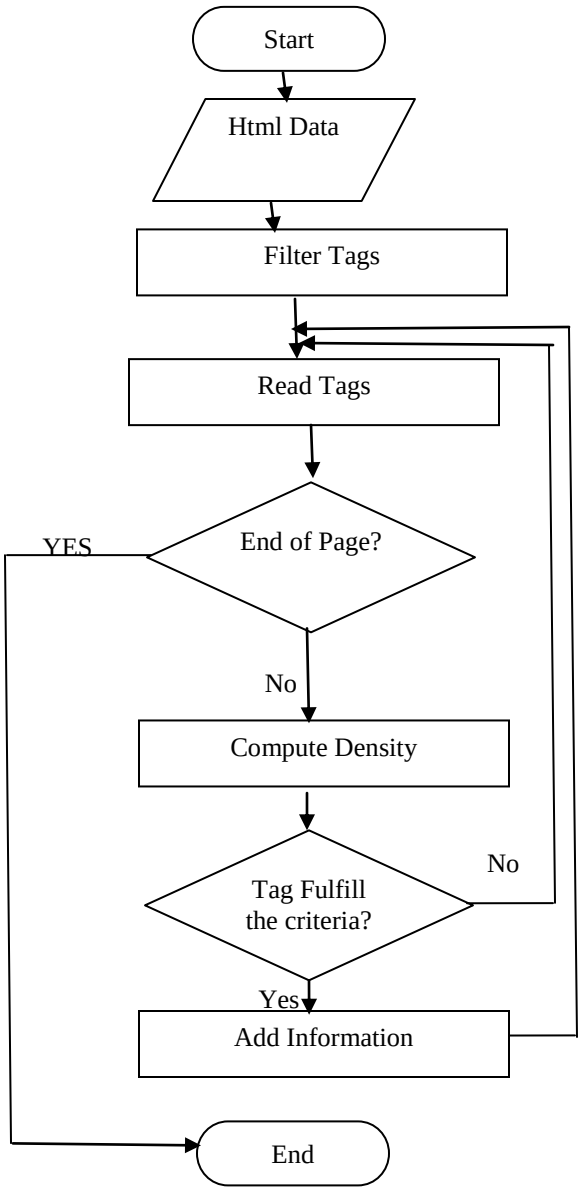


Figure 2 Parse Html Flow Chart

Before parsing the html data the Comment, script, input button, link, style, select, embed, object, img and iframe tags are removed in the tag filtering step because these tags are likely to have noise and do not contain the data we are interested in. Each tag's density will be the total number of words under that tag. So tags with the density higher than particular threshold (15 words per tag) will make output content block. But there may be possibility that tag with lesser density also have the some important information. To remedy this we make a list of keywords from the title of the page and we check if two occurrence of keywords is found we add it the output block. When making keywords helping verb like has, have and any other grammatical words that are irrelevant should be ignored.

For example the html page shown in figure 3(a) and 3(b) has the following title.

<Title>Mars Mission can see India emerge major power in science and technology </title>

So the keyword for this will be

Keywords = {mars, mission, india, emerge, major, power, science, technology}

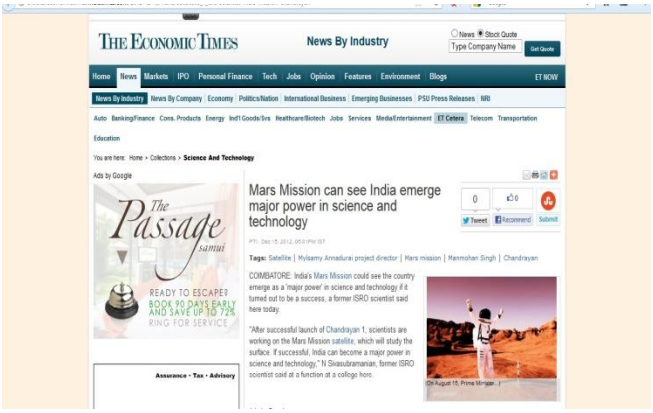


Figure 3(a)

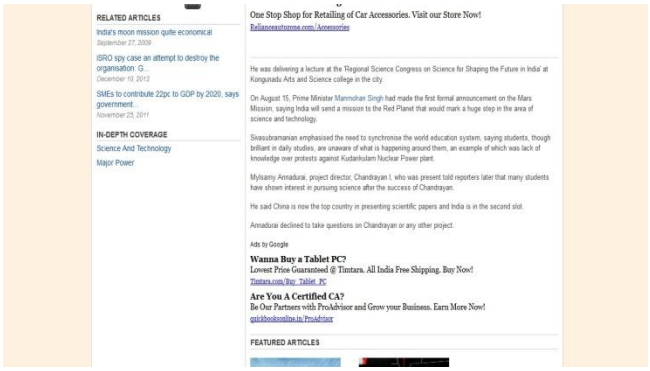
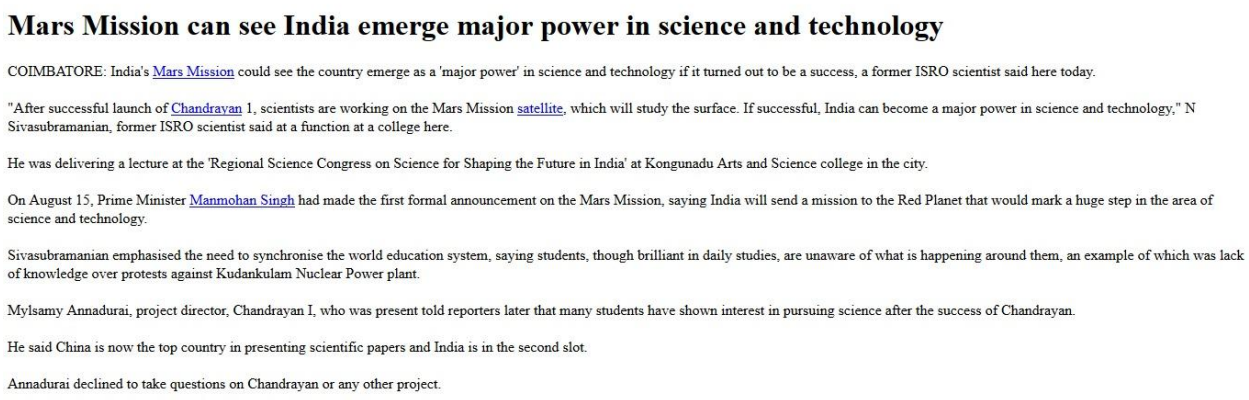


Figure 3(b)

The output of the method is shown in figure 4. As you can see from the output all the noises has been removed and only the main block is shown.

3. RESULT EVALUATION

The result will be evaluated on the basis of three factor completeness, correctness and quality. Completeness tells us how complete our approach is i.e. how much information we are extracting from the main content of the web page. Correctness tells us how correct those information are. Quality tells us how complete and correct



Mars Mission can see India emerge major power in science and technology

COIMBATORE: India's Mars Mission could see the country emerge as a 'major power' in science and technology if it turned out to be a success, a former ISRO scientist said here today.

"After successful launch of Chandrayan 1, scientists are working on the Mars Mission satellite, which will study the surface. If successful, India can become a major power in science and technology," N Sivasubramanian, former ISRO scientist said at a function at a college here.

He was delivering a lecture at the 'Regional Science Congress on Science for Shaping the Future in India' at Kongunadu Arts and Science college in the city.

On August 15, Prime Minister Manmohan Singh had made the first formal announcement on the Mars Mission, saying India will send a mission to the Red Planet that would mark a huge step in the area of science and technology.

Sivasubramanian emphasised the need to synchronise the world education system, saying students, though brilliant in daily studies, are unaware of what is happening around them, an example of which was lack of knowledge over protests against Kudankulam Nuclear Power plant.

Mylsamy Annadurai, project director, Chandrayan I, who was present told reporters later that many students have shown interest in pursuing science after the success of Chandrayan.

He said China is now the top country in presenting scientific papers and India is in the second slot.

Annadurai declined to take questions on Chandrayan or any other project.

Figure 4 Output

our approach is i.e. more complete and more correct method means better quality.

Completeness

$$\text{Completeness} = TP/(TP+FN)$$

$$\text{Completeness} \in [0; 1]$$

Correctness

$$\text{Correctness} = TP/(TP+FP)$$

$$\text{Correctness} \in [0; 1]$$

Quality

$$\text{Quality} = TP/(TP+FP+FN)$$

$$\text{Quality} \in [0; 1]$$

TP = True positives the number of tags that are actually informative i.e. not a noise

FP = The number of tags that are noise but counted as informative tags.

FN = The number of tags that are informative but counted as noise

The result from various websites is shown in the table 1.

Table 1 Results

URL	Completeness	Correctness	Quality
http://articles.economictimes.indiatimes.com	1	1	1
http://www.indianexpress.com/	0.93	0.88	0.83
http://www.thehindubusinessline.com	0.92	0.86	0.80
http://www.business-standard.com	0.92	0.92	0.85
http://www.thehindu.com	0.88	0.88	75
http://www.guardian.co.uk	0.86	0.95	0.83
http://www.dnaindia.com	0.85	0.81	0.71
http://www.firstpost.com	0.93	0.87	0.81
http://www.dailymail.co.uk	0.86	0.86	0.76
http://ibnlive.in.com	0.93	0.83	0.78

4. REFERENCES

[1] Shin, Kwangcheol, and Geun Sik Jo. "Catch Crawler: Automatic Web Information Extractor Using Style Sheet." Semantic Computing and Applications, 2008. IWSCA'08. IEEE International Workshop on. IEEE, 2008.

[2] Sun, Fei, Dandan Song, and Lejian Liao. "Dom based content extraction via text density." SIGIR. Vol. 11. 2011.

[3] Asfia, Mohsen, Mir Mohsen Pedram, and Amir Masoud Rahmani. "Main Content Extraction from Detailed Web Pages." International Journal of Computer Applications IJCA 4.11 (2010): 18-21.

[4] Downey, Doug, et al. "Learning text patterns for web information extraction and assessment." AAAI-04 workshop on adaptive text extraction and mining. 2004.

[5] Yi, Lan, and Bing Liu. "Web page cleaning for web mining through feature weighting." International joint conference on artificial intelligence. Vol. 18. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003.