International Journal of Multidisciplinary Research and Development 2015; 2(4): 10-12

International Journal of Multidisciplinary Research and Development

IJMRD 2015; 2(4): 10-12
www.allsubjectjournal.com
Received: 22-03-2015
Accepted: 03-04-2015
e-ISSN: 2349-4182
p-ISSN: 2349-5979
Impact Factor: 3.762

Priyanka Gangurde
Computer Department, PREC Loni, Maharashtra, India

Dipali Rakh
Dipali T. Rakh, Computer Department, PREC Loni, Maharashtra, India

Sushant Valvi
Sushant K. Valvi, Computer Department, PREC Loni, Maharashtra, India

Correspondence:
Priyanka Gangurde
Computer Department, PREC Loni, Maharashtra, India

Online news web text extraction based on modified maximum subsequence segmentation

Priyanka Gangurde, Dipali Rakh, Sushant Valvi

Abstract
In daily lives we use the web as main information source. We search the online news on the web pages. A huge amount of advertisements and external links on the news web page which is complicated to extract the news from the original HTML document. News in web pages contain a lot of extra contents like advertisements, headers, footers, external links and navigation bar which is not useful for the users. Redundant and irrelevant information is distributed and mixed in whole page, it is hard for user to automatically identify the useful information in the page. This not only increases the cost for search on Web pages, but also it is difficult for users of small display devices. In this paper, we proposed the maximum subsequence segmentation algorithm for extract the news from the web page and convert it into Multilanguage. We get accurate result by using maximum subsequence segmentation algorithm.

Keywords: News reading, Communication education, Automatic news content extraction, Multilanguage learning etc.

1. Introduction
The Web information has become increasingly in different way. Utilize the Web information good, people use the latest technology, which can effectively organize and use online information. Lot of noise contains in web information. For example, the scripts are added for increasing user's interactivity; the navigation links are added for facilitating the user's browse; and advertising links are added for commercial factors. Many content of the information on the web is found in articles from online news outlets, review collections, magazines, and other resources. Extraction of this content from the original HTML document is complicated and difficult by the large amount of less informative and typically unrelated material such as navigation menus, forms, user comments, and ads. These information affect the web information extraction efficiency and lead to the accuracy decline. Pre-processing stage of web information processing is quickly and accurately identifying the text of the page has become an essential. One best advantage of online news reading is that it can fast distribute to the world wide and distribute to people in a very short time. Some programs and courses have started to adopt the online news as the full course. It is easy for the user for the reading of news as their time saves. User should read a news and try to re-write a fused news report or try to organize the aspects of different perspective views. However, building news data for human is very time-consuming since the news extraction is manually collected and stored by the introduction of information extraction technology.
We proposed the maximum subsequence segmentation algorithm for extract the news from the web page and convert it into Multilanguage. As compare to previous algorithms the results are more accurate. We apply the maximum subsequence segmentation algorithm where we use the negative and positive scores for the extraction of the HTML document. The problems of learner while reading the newspaper reduced by Maximum subsequence segmentation algorithm. Here we convert the news in Multilanguage as the main problem of the learner is the language where the many news are in the different languages hence we convert the news into the Multilanguage as they want. We modify the maximum subsequence segmentation algorithm and we get the more accuracy than previous. We get accurate result by using maximum subsequence segmentation algorithm.

2. History:
Literature showed that, according to the label statistical information, such as the number of characters, the web page is building into a DOM tree based on node, links to extract web text. The method is very simple, and does not depend on a specific kind of data, also

achieves better effect. The Web page tends to be increasingly complicated. Many web pages contain a huge amount of noise blocks and not important information which disturb the text feature information statistics; thus, the noise blocks make the web text owning some noise or missing part of the text information. A fault is that the HTML's standard needs is higher, and a lot of memory for element node object is taken up. This algorithm depends on many text characteristic attributes, and calculation multifarious, its adaptability is limited. It differs from the extraction of the DOM tree, the algorithm is based on the label and text distribution. The algorithm ignores the grammatical structure of web page source code, but accords to the distribution of the label to distinguish between text and non-text. Due to the analysis of building DOM tree by the web page source code structure, the efficiency is very high, though it is limited by the web page position.

The method is very simple, and does not depend on a specific kind of data, also achieves better effect. The Web page tends to be increasingly complicated. Some web pages contain a large amount of noise blocks which disturb the text feature information statistics; thus, the noise blocks make the web text owning some noise or missing part of the text information.

Literature Wang Li, LiuZongtian, WangYanHua[1] showed that, through the text feature attributes the particle swarm algorithm to optimize the feature weights and threshold value, so as to determine the text information in the block. But this algorithm depends on many text characteristic attributes, and calculation multifarious, its adaptability is limited.

Literature S.Chakrabarti, M. Berg, B. Dom[3] showed that, differ from the extraction of the DOM tree, the algorithm is based on the text and label distribution. The algorithm ignores the grammatical structure of web page source code, but accords to the distribution of the label to distinguish between text and non-text. Due to the analysis of building DOM tree by the web page source code structure, the efficiency is very high, though it is limited by the web page position. Extracting article text, as with cleaning HTML documents in general, is sometimes dismissed as easy or even trivial.

Barzilay, McKeown, K.R., Evans, Hatzivassiloglou,V., Klavans, J.L., Nenkova, A. [4] spend a single sentence describe how they obtained the desired text from a page with a simple heuristic. Unfortunately, this sentiment is antiquated: over time, page layout has become much more complex as style has evolved from the unadorned, flat documents of the early web to basic frames and nested structures, with navigation panels, sidebars, inserts, headers, ads, etc., and the rudimentary techniques that worked years ago are inadequate today. Unsurprisingly, their heuristic, as our first baseline, proves to be quite inaccurate. The first step in an information extraction pipeline & cleaning errors propagate through the system and harm overall performance. Accordingly, the most common approach has been to hand-code rules, often as regular expressions; these can be very accurate, but are highly labor intensive and easily broken by changes to a page's structure. This has led to interest to finding a better solution, as exemplified by the recent Clean Eval shared task.

VIPS (vision-based page segmentation) the Method is put forward by the Microsoft research Asia. By making use of visual information in HTML, such as the background color, the font color and size, the border, the logical block and the logical block distance, etc., VIPS algorithm identifies the page same block, and then combining with DOM tree structure to realize page web text extraction. Although VIPS algorithm can achieve good results .It has some drawbacks, when extracting pages visual information, it over relies on the browser kernel, time-consuming, and the algorithm is relatively complicated. Yu et al[2]. uses the vision-based content structure of Web pages to extract relevant blocks. It also uses a hierarchical structure method indicating the different views of a Web page. VIPS and the hierarchical methods DOM-based extraction methods are use for tagging structure and tag semantics to recognize block boundaries and block semantics, but they do not consider context same, which may be useful and redundant.

VIPS can extract article text with a good accuracy, but creating a mechanism to choose among the dozens or hundreds of nodes is a difficult task in itself. Additionally, VIPS must partially render a page to analyze it: while it does not need to actually draw DOM elements to the screen, it still necessary to parse the page, build a DOM tree, and determine element layout. If external style sheets are used, these should be retrieved. Consequently, compared to other techniques, VIPS is resource-intensive, which impedes its scalability to larger numbers of documents. There are different methods like VIPS, attempt to identify generally interesting, informative portions of a document rather than specific fields.

Chen, Ye & Li split pages on certain tags into blocks and cluster them based upon style and position; similar clusters among different pages are then identified as part of the template. Yi &Liu and Yi, Liu & Li create a compressed structure tree and a site style tree that identify DOM nodes with similar style across pages as uninformative. Ma et al.opt for a simpler approach, dividing a page according to <TABLE> tags and looking for frequently repeated segments. Lin &Ho also partition a page with <TABLE> tags, but uses a more sophisticated method whereby redundant blocks are identified using an entropy measure over a set of word-based features. However, these template detection approaches tend have three major weaknesses: as with Road Runner, multiple pages known to have the same template are required, and they often incorrectly assume that certain tags always delineate segments of interest or that uninteresting segments are largely repeated across pages (clearly not the case for "junk" like varying text ads or lists of related articles).

3. Working diagram:

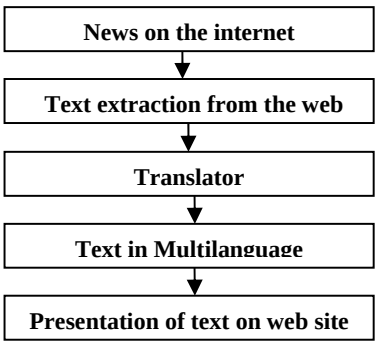

Fig 1: working diagram

Users search the news on web. Extraction of the news is the difficult from the HTML document. So we developed the algorithm, which extract text automatically from the web pages that is extraction of the news from the web site by

using maximum subsequence segmentation algorithm. Translator is used for translate the news in Multi languages such as Marathi, Tamil, Hindi, Urdu and so on. We proposed the new technique of the language translator in that there is no need of text copy and paste. We add the button over the web page and after clicking on the button the translator directly gives the result. Translator gives the result very quickly on web as compare to the other.

4. Algorithm:

Given a sequence R= (R_1, R_2, R_3…..R_n), where $R_{Di}\in$,
Maximum subsequence of R is a sequence T= (R_p, R_q…..R_n)
Where, $1 \leq p \leq q <$n and p, q is given by:

$$(p,q) = \arg max_{(a,b)} \sum_{i=a}^{b} Ri$$

T is constrained to be a substring since we specify that all elements are consecutive, and it would be accurately the maximum substring. The sequence may be contain more than one maximum subsequence and the problem is trivial when all elements of the sequence are no-negative. The Maximum subsequence segmentation algorithm extract the text by calculating the positive and negative scores of the html document. We assign the negative score to the html tags and positive for the text. We gives the -1 to the html tags and the +1 to the text. After that we calculate the sum of positive score and negative score separately and then add both scores. By this calculation we get better result for extraction of the html document.

The working of algorithm is given as follows:

Input: Sequence of calculated score
$R = (R_1, R_2,\ldots,R_n)$
Process:
tagsum=0;
textsum=0
for i=1 to n :
while Ri=-1 do
tagsum -= 1;
i++;
end while
pos=i;
while Ri=1 do
textsum += 1;
i++;
end while
Sum=tagsum+textsum;
If sum is positive then
Get text from pos to i
End if
End for
Output:
Return Extracted article text

4.1 Results:

In order to evaluate the accuracy of the algorithm we proposed in this paper. In the process of news web text extraction, we determines the accuracy of the text extraction During extraction of text , the text information not lost, and even no small value will take noise information included. We converted the extracted text into Multilanguage. The result of the text extracted from the web information is consistent with human observation. Extraction result only include all text information, not contains a small amount of the non-text information.

5. Conclusions

This paper puts forward algorithm which is based on the maximum subsequence segmentation. The extracting web text problem is gradually changed into the largest continuous subsequence sum. In the complicated situation, in order to extract web text, we define the concept of web candidate text and text degree. Result show that, news content extraction techniques while removing all non-content areas which provide a clean page for learners. The news convert into multiple languages. Automatically developing a news reading platform for communication education is a new research topic in recent years. This problem involves integrating news content extraction techniques while removing all non-content areas which provide a clean page for learners. The experimental result showed that this method did not only achieve the best result in accuracy in comparison to previous studies, but also save more time for reading news. In the future, we investigate the use of our news reading platform to different language learning applications.

Acknowledgement

First and foremost, I would like to thank my guide, Prof.Vikhe P.B., for his guidance and support. I will forever remain grateful for the constant support and guidance extended by guide, in making this report. Through our many discussions, he helped me to form and solidify ideas. The invaluable discussions I had with him, the penetrating questions he has put to me and the constant motivation, has all led to the development of this project.

I wish to express my sincere thanks to the Head of department, Prof. Jondhale S.D. also the departmental staff members for their support. I would also like to thank to my friends for listening to my ideas, asking questions and providing feedback and suggestions for improving my ideas.

References

1. Wang Li,Liu Zongtian,Wang YanHua the extraction of web page text based on content similarity. in computer Engineering,2010
2. D. Cai, S. Yu, J.-R. Wen and W.-Y. Ma.”VIPS: a vision-based page segmentation algorithm” in Microsoft Technical Report MSR-TR-2003-79, 2003.
3. S. Gupta, G. Kaiser, D. Neistadt, P. Grimm, DOM-based Content Extraction of HTML Documents, in 12th World Wide Web Conference, 2003.
4. McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B. and Sigelman, S. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. HLT 2002.