DOM Tree based Approach for Extraction of Content from Web News Page

Snehal Dilip Khade
Department of Computer Engineering
K.J. Somaiya College of Engineering
Vidya-Vihar, Mumbai, India
snehalkhade9@gmail.com

Prof. Jyothi Rao
Department of Computer Engineering
K.J. Somaiya College of Engineering
Vidya-Vihar, Mumbai, India
jyothimrao2004@gmail.com

Prof. Manish Potey
H.O.D. Department of Computer Engineering
K.J. Somaiya College of Engineering
Vidya-Vihar, Mumbai, India
manishpotey@gmail.com

Abstract— With the huge amount of data on the World Wide Web, information extraction has become an increasingly important technology to help users locate desired Web information. Due to the heterogeneous nature of Web content, it is difficult to design a general Web information extracting approach that fits all application domains. When building a system for searching or mining Web content, a first task is extracting the main content and removing extraneous data such as navigation menus, functional and design elements, and commercial advertisements. It is very important to filter out such noise from web pages like web new pages. Also, when showing Web news pages on small screens like mobile phones or sending text to screen readers that translate the text to a more appropriate format like text-to-speech for visually impaired people, the content extraction operation is very valuable. Content extraction is defined as the process of determining those parts of an HTML document that represent the main textual content. The problem, however, is to find a solution that is generic which is portable to many types of Web news pages, accurate that finds all important content in a precise way and efficient where often a large number of Web pages are processed . An approach is designed which searches for relevant web pages using a web crawler then using a DOM tree based approach extracts the content from web news page by filtering out noise and the information retrieval agent extracts the key paragraph from the extracted content.
Keywords-information extraction; Web Crawler, Document Object model; Web content extraction; information retrieval agent;

I. INTRODUCTION

With the development of the Internet, the Web is becoming the largest data repository ever available in the history of humankind. The Web keeps growing and huge amount of new information are being posted on it continuously. Weekly, tens or hundreds of Megabytes of news stories can be added easily to the news archive of any news sources online. At the same time containing some influencing knowledge, this news archive may also be holding many uninteresting or trivial news. The influencing knowledge is desired but reading the news archive is rather a daunting task that will takes a lot of time and effort.

Major efforts have been made in order to provide efficient access to relevant information within the web pages. Today's Web pages are commonly made up of more than merely one cohesive block of information. So extracting exact information content becomes difficult. For instance, news pages from popular media channels such as Financial Times or Washington Post consist of no more than 30%-50%of textual news, next to advertisements, link lists to related articles, disclaimer information and so forth. Web pages are often decorated with extraneous information which includes navigation bars, branding banners, JavaScript, links to related news, copyright claimant and advertisements, copyright notices, privacy policies, comments, headers, footers. This kind of information distracts users from actual content they are interested in and also degrade the performance of information retrieval applications. Therefore, a method to identify and extract main and relevant content is needed to alleviate this problem. Efficiently extracting high-quality content from Web page is crucial for many Web applications such as information retrieval,

IJKRCE

automatic text categorization, topic tracking, machine translation, abstract summary, helping end users to access the Web easily over constrained devices like PDAs and cellular phones. Most previous works rely on the template of the web sites. When information like news needs to be extracted from different sites, it must create a template for every site which will spend much time and huge cost. Many algorithms have been proposed to identify relevant pages but have certain disadvantages due to the heterogeneous nature of a web pages. A generic approach is needed that extracts real content from Web pages in an unsupervised fashion.The proposed work develops an approach to extract web news content from a number of Web news pages, and these pages come from a great many of heterogeneous Web news sites. It does not gives just the links as search engines do. Initially it extracts the relevant web pages and by using DOM tree approach it extracts the content from these pages after which using a information retrieval agent only the key paragraph from the whole content is presented to the user.

Contents of our paper is as following. In section 2, we introduce briefly relevant work done. Section 3 defines proposed model for In section 3, we explain proposed DOM tree based approach for extraction of content from web new page. Section 4 gives the conclusion.

II. RELATED WORK

To identify the actual content in the Web page is a relatively easy task for a human who can do it just by visual inspection, however it is a hard problem for a computer.

There have been many approaches existed to extract content from Web news page like Wrapper induction, using web techniques and techniques based on statistics. A wrapper[5] can be generated by wrapper induction system for content extraction from Web news page such as . However one wrapper is usually being generated for only one information source. Since there are so many heterogeneous web sources, it is not practical to build wrappers for each web source .Therefore, this class of approaches is not fit for our task. Here are some classical works: Supervised approaches WIEN,STALKER and unsupervised approaches Dela Road Runner, and EXALG .

Some approaches use some techniques of Web mining, such as classification and clustering[2][3], to extract content from Web page These approaches can improve the accuracy of extraction. However most of them need human interventions, and the complexity of the underlying algorithms is not low, so this class of approaches has limited ability for scalable extraction. The problem of identifying content from a Web page is treated as a sequence labeling problem. The content of a Web page is identified by using a Conditional Random Field sequence labeling model. In traditional hierarchical clustering techniques are used to extract the desired content from Web sites.

Some approaches extract content from Web page based on statistics[11]. These approaches can usually perform the extraction in an unsupervised fashion, which is crucial for the task. However most of them rely on some weights or thresholds that are usually determined by some empirical experiments. It is difficult to find one set of weights or thresholds to satisfy all news pages coming from so many heterogeneous sources.

Thus, wrapper is not suitable for many heterogeneous web pages, classification and clustering needs human intervention and statistics techniques uses set of weights or thresholds to satisfy all news pages coming from so many heterogeneous sources.

The proposed approach resolves all the problems specifically human intervention. It uses DOM tree approach which supports navigating and modifying and then extracts the content by filtering noise. The Document Object Model is an application programming interface for valid HTML and well-formed XML documents. It defines the logical structure of documents and the way a document is accessed and manipulated. In the DOM specification, the term "document" is used in the broad sense - increasingly, XML is being used as a way of representing many different kinds of information that may be stored in diverse systems, and much of this would traditionally be seen as data rather than as documents. Nevertheless, XML presents this data as documents, and the DOM may be used to manage this data.

With the Document Object Model, programmers can build documents, navigate their structure, and add, modify, or delete elements and content. Anything found in an HTML or XML document can be accessed, changed, deleted, or added using the Document Object Model, with a few exceptions - in particular, the DOM interfaces for the XML internal and external subsets have not yet been specified. With this features it becomes easier to extract the useful content from a large number of heterogeneous web pages. Also this approach is generic, that is, it is not bounded to any templates and thresholds. Also the proposed work has information retrieval agent[7] based on meaningful term's frequency and the key word distribution characteristics in a document.

III. OUR APPROACH

The user specifies his required information to the system. The web crawler takes its seed URL and searches for the relevant pages. Then DOM tree is generated of those pages. Now the irrelevant contents

IJKRCE

like advertisements, link lists to related articles, disclaimer information, user comments, navigational menus, headers, footers, copyright notices, privacy policies, are removed using algorithms Joint-para and Extract-news which uses the DOM tree structure. Then the key paragraph from the extracted content is extracted using information retrieval agent.

The following flow will explain about the methodology used:

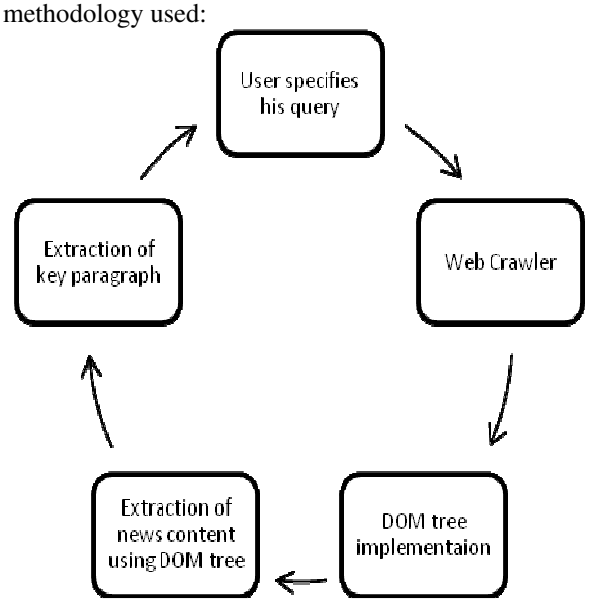

Figure 1. Flow of the proposed approach

The flow from web crawler to DOM tree implementation plays a crucial part in the proposed work.

A. Web Crawler

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. Therefore, any web crawler needs to keep track both of the URLs that are to be downloaded, as well as those that have already been downloaded (to avoid unintentionally downloading the same page repeatedly). This is achieved using data structure on disk. Virtually every modern web crawler splits the crawl state into two major data structures: One data structure for maintaining the set of URLs that have been discovered (whether downloaded or not), and a second data structure for maintaining the set of URLs that have yet to be downloaded. The first data structure (sometimes called the "URL-seen test" or the "duplicated URL eliminator") must support set addition and set membership testing, while the second data structure(usually called the frontier) must support adding URLs, and selecting a URL to fetch next. Thus the output of a web crawler will be a set of relevant web pages.

B. Document Object Model

The relevant pages given out by the web crawler are represented in a form of DOM tree

HTML DOM is in a tree structure, usually called an HTML DOM tree. Figure 2 illustrates a simple HTML document and its corresponding DOM tree. We are interested only in the <BODY> node and its offspring. In this example, <BODY> node has three children: element nodes <B> and <I>, and text node #and. Element node <B> has a text node child #bold, and element node <I> has a text node #italic. Following the DOM convention, we use <> to indicate element node, and use # to indicate text node.

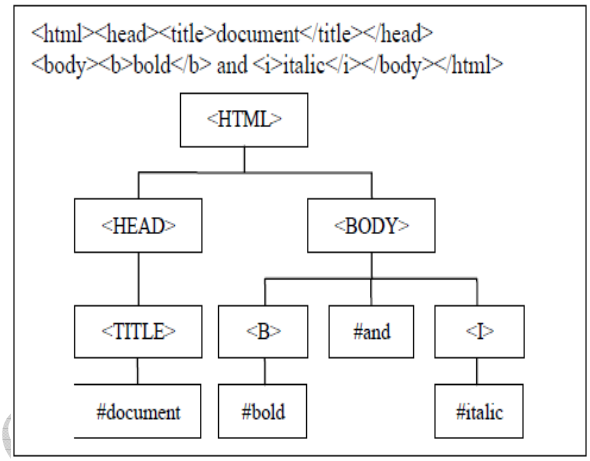

Figure 2. Simple HTML document and its corresponding DOM tree.

C. Extraction of News Content using DOM tree

After representing the relevant web news pages in a form of DOM tree structure, the following two algorithms are used to extract the content by filtering out noise.

1.Joint-para Algorithm
2. Extract-news

Prior to starting with algorithms there are certain following observations as follows

1)There is such node that the entire content of news is wrapped in it with its subtrees, and any subtree of it cannot wrap the entire content of news. Such node is called as summary-node. That is, the summary-node with its subtrees is the minimal tree which contains the entire content of news. Note that there may be some noise embedded within some subtrees of the summary-node. The summary-node of Figure 2 is in a star in Figure 3. The tag-name of the summary-node is .

2) There is such node that it is the descendant of the summary-node, and it is the father of a text node, and a part of or entire content of news is wrapped in it with its subtrees. Such node is called as snippet-node.

3) The node which satisfies either of the conditions is called as big-node:

a) The node is the father of a text-node and the tag-name of the node is not <script>or <style>;
b) The tag-name of the node is <p>or
or <h1>or <h2>or <h3>or <h4>or <h5>or <h6>oror or
or <b>or <i>or <tt>or <font>.
The content of news is usually broken into many small pieces by these nodes.
4) One set of nodes that satisfies the conditions is called as text-node-set:
a) The nodes in the set are all big-nodes, and they are at the same level in the DOM tree and they are adjacent;
5) The nodes in the set together wrap a part of or entire content of news or noise, and the text wrapped by the set is called as a text-para.

6)The number of period and comma in a piece of text is called as punc-num. The punc-num in the text wrapped bya node is called as node-punc-num.

Algorithms to extract web news content

1. Joint-para Algorithm
If there is a long piece of noise, it is easy to wrongly regard the piece of noise as the start point of backtracking. To guarantee finding a correct start point of backtracking, the algorithm of Joint-para will merge short pieces of text. At the same time, some noise may be embedded within some subtrees, so Joint-para needs to prude some noisy nodes during merging.
Given: Input:A DOM tree
(entire text of news is broken into many short pieces by some nodes such as <p>and
.)
1) Trace the tree breadth wise to find the big node B1
(The node is the father of a text-node) &&(the tag-name of the node is not <script>or <style>;)&& (The tag-name of the node is <p>or
or <h1>or <h2>or <h3>or <h4>or <h5>
or <h6>or or or
or <b>or <i>or <tt>or <font>.)
2)Big node B1 checks its adjacent nodes to find a big node and thus it forms a text-node-set
One set of nodes that satisfies the conditions is called as text-node-set. The nodes in the set are all big-nodes, and they are at the same level in the DOM tree and they are adjacent
3) The nodes in the set together wrap a part of or entire content of news or noise, and the text wrapped by theset is called as a text-para
4) Compute the punc-num of the text-para.
(The number of period and comma in a piece of text is called as punc-num. The punc-num
in the text wrapped bya node is called as node-punc-num.)
5) If punc-num is 0,
a) the text-para will be regarded as noise and will not be output.
b)all the nodes that together wrap the noise piece are pruned.
6) If the punc-num is not 0,
a) the text-para will be output
7) End

2. Extract News Algorithm
The heuristics to detect when to stop backtracking is from such observation: When backtracking from node1 to node2,if the content of news wrapped by node2 is more than node1,node1 must not be the summary-node, and the node-puncnum of node2 must be more than node1. If node1 is the summary-node, there are two cases as following:
1) The information wrapped by node2 is equal to node1,so the node-punc-num of node2 must be equal tonode1;
2) There is more noise wrapped in node2 than node1,and the extra noise does not contain any period and comma, so the node-punc-num of node2 must be equal to node1.
Algorithm of Extract-news:
1) Perform Joint-para for each big-node to get all text paras.
2)Then select randomly one node from the text-nodeset that wraps the longest text-para.
The selected node is regarded one snippet-node.
3)Then backtracking starts from the snippet-node.
When backtracking from node1 to node2,calculate the node-punc-num of node1 and that
of node2 respectively. Then use the node-punc-num of node2 minus the node-punc-num of node1, and get the difference that is called as distance. Thus on the way of backtracking,
a sequence of distance can be obtained.
4)The process of backtracking stops at the following condition:
The distance appears 0 for the first time. For the distance 0, the child node is regarded as
the summary-node.
5) At last, the content wrapped by the summary-node is extracted as the entire content of news.
6)End

D. Extraction of Key paragraph from the extract content

The extracted content from a web news page now further undergoes filtration where only the relevant paragraph is extracted from the whole content.
The information retrieval agent using a web client's information retrieval request query extracts key paragraph with frequency and distribution using the keywords of the client, and then the agent constructs profile of the documents with the keywords. A document is composed of many terms and important words are spread out documents. In those documents, we want to get a meaningful or worth terms and the word indexed. It is important for efficient information retrieval or knowledge discovery that indexing is defined very well by appropriate terms about all documents. We propose a new method for key paragraph extraction, which compute term frequency

JKRCE

and key word distribution of each term selected by using stemming, filtering stop-lists, synonym for search meaningful terms in a document. Using the extracted paragraph with frequency and distribution, we construct profile with the index, key paragraph of the document. And then we can search many documents or knowledge easily using the profile for information retrieval and browsing document. After we extract worth terms using stemming, filtering stop lists, synonym, etc. we define location of the terms in document. The criterion of term's location can be defined various type, document line or sentence, etc.

Our proposed indexing algorithm is as following.

ti : ith meaningful term in a document
fi : frequency of ith term in a document
lij : if ith term appears in jth location of the document, 1, otherwise 0
di : if ith term's frequency is greater than a criterion, 1, otherwise 0

1. Extract worth terms in each of the paragraph of a document using stemming, filtering stop-lists, synonym, etc. This processing is scanning a document for searching meaningful terms(t_i). And examine the term's location of document, where location can be defined line or sentence.
2. When extract meaningful term, compute terms frequency(f_i), location of document(l_{ij}) respectively.
3. Create table about each terms and frequency. And compute summation of term frequency in each location of document.

After extract meaningful terms, some terms can be eliminated by criterion of frequency(D). For example, when a term is appear less than three times, the term is treated as worthless term. When a term t_i is defined worth term or not, we use denotation $d_i = \{0,1\}$, meaningful 1, worthless 0. When we need to extract more keyword, we can change criterion of frequency(D).

4. The region where summation of frequency is greater than a criterion is extracted key paragraph and terms in the key paragraph can be defined as keywords.

Summation of frequency(s_j) is computed as

$S_j = \sum(d_i \times l_{ij})$

where $d_i = 0$ or 1 and $l_{ij} = 0$ or 1.

We can select region as key paragraph where summation of frequency is greater than a criterion on paragraph. And terms in the selected paragraph are defined keywords.

Thus the output of the above method will give the most relevant content to the user.

IV CONCLUSION

In this paper we have proposed a method which extracts the most relevant content and presents it to the user. The user specifies his query, accordingly a web crawler extracts and downloads relevant web pages .Using DOM tree approach contents of the web pages are extracted by filtering out noise. By using information retrieval agent the most relevant content from the whole web page is presented to the user. This approach is a generic one and works for heterogeneous pages.

With the Document Object Model, programmers can build documents, navigate their structure, and add, modify, or delete elements and content. With this features it becomes easier to extract the useful content from a large number of heterogeneous web pages. Also it is efficient, it not just gives the links to the relevant pages like search engines do, but gives the most relevant content. It has its applications in showing news on small screens like mobile phones or sending text to screen readers that translate the text to a more appropriate format like text-to-speech for visually impaired people, here the content extraction operation is very valuable.

In future this approach will be used in information retrieval, automatic text categorization topic tracking, machine translation, abstract summary. It can provide conceptual views of document collections and has important applications in the real world.

IJKRCE

REFERENCES

[1] Yan Guo ,Huifeng Tang , Linhai Song , Yu Wang and Guodong Ding, " ECON: An Approach to Extract Content from Web News Page", IEEE 2010, 978-0-7695-4012-2/10.

[2] J. Prasad and A. Paepcke, "Coreex: content extraction from online news articles," in CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management New York, NY, USA: ACM, 2008, pp. 1391–1392.

[3] J. Gibson, B. Wellner, and S. Lubar, "Adaptive web-page content identification," in Proceedings of the 9th annual ACM international workshop on Web information and data management. ACM New York, NY, USA, 2007, pp. 105– 112.

[4] S. Gupta, G. E. Kaiser, P. Grimm, M. F. Chiang, and J. Starren, "Automating content extraction of html documents,"World Wide Web, vol. 8, no. 2, pp. 179–224, 2005.

[5] N. Kushmerick, "Wrapper induction for information extraction," Ph.D. dissertation, 1997, chairperson-Daniel S. Weld.

[6] Christopher Olston and Marc Najork "Web Crawling" Foundations and Trends in Information Retrieval Vol. 4, No. 3 (2010) 175–246, 2010 C.

[7] Jae-Woo LEE "A Model for Information Retrieval Agent System Based on Keywords

Distribution" International Conference on Multimedia and Ubiquitous Engineering(MUE'07), IEEE 2007 0-7695-2777-9/07

[8] Marti A. Hearst andXerox PARC "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages" Association for Computational Linguistics, 2007

[9] D. C. Reis, P. B. Golgher, A. S. Silva, and A. F. Laender, "Automatic web news extraction using tree edit distance," in WWW '04: Proceedings of the 13th international conference
on World Wide Web. New York, NY, USA: ACM, 2004, pp. 502–511.

[10] K. McKeown, R. Barzilay, J. Chen, D. Elson, D. Evans, J. Klavans, A. Nenkova, B. Schiffman, and S. Sigelman, "Columbia's newsblaster: new features and future directions,"in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations-Volume 4. Association for Computational Linguistics Morristown, NJ, USA, 2003, pp. 15–16.

[11] S.-H. Lin and J.-M. Ho, "Discovering informative content
blocks from web documents," in KDD '02: Proceedings of the
eighth ACM SIGKDD international conference on Knowledge
discovery and data mining. New York, NY, USA: ACM,2002, pp. 588–593.

JIKRCE