Efficient Content Extraction Using Hybrid Technique

G. Naveen Sundar, Sheba Gaikwad
Dept. of Computer Science, Karunya University, India

Abstract
Content extraction is the process of identifying the main content or removing the additional contents. The main problem in extracting the content from the web page is the newer architecture of web pages and the diversity in the structure of web pages. Many content extraction strategies are based on DOM tree representation, feature extraction or tag ratios of HTML web page and extracting useful content from it. This paper describes a comparative study on various content extraction algorithms.

Keywords
Data Mining, Information Extraction, Content Extraction, HTML, Open Source Intelligence, Information Filtering

I. Introduction
The information available today on web is tremendous and comes with greater challenges. There is a need of proper tool namely capable of automatically processing the online information to deal with such information overload effectively. The difficulty in extraction is that the data is mixed together with formatting code, advertisement, or web page specific information such as navigation links, in a way that the web page is human friendly, but the actual content is not machine friendly this complicates the task of automatic content extraction and processing. Content Extraction is useful for visually impaired and blind. Natural Language Processing (NLP) and Information Retrieval (IR) can also benefit from content [4]. Currently, most of the NLP based IR applications require writing specialized extractors for each of the web domain [4]. Extraction is particularly useful for Open Source Intelligence (OSInt), which relies on automatic extracting information from web pages then processing the extracted information to gain knowledge.

A typical web page contains a title banner, list of links in right or left or both for site navigation and advertisements, a footer containing copyright statements, disclaimers or navigational links [9]. Mostly, meaningful content lies at the centre of the page. The design of web page is not standard for all web pages, consequently, a more robust and flexible content extraction tool is essential. Recent web pages have a cleaner architecture. They provide separation among visual presentation, real content and navigation links, in that way that web page is human friendly, but the actual content is not machine friendly this complicates the task of automatic content extraction and processing. Content Extraction is useful for visually impaired and blind. Natural Language Processing (NLP) and Information Retrieval (IR) can also benefit from content [4]. Currently, most of the NLP based IR applications require writing specialized extractors for each of the web domain [4]. Extraction is particularly useful for Open Source Intelligence (OSInt), which relies on automatic extracting information from web pages then processing the extracted information to gain knowledge.

B. Discovering Informative Content Blocks [5]
The system discovers informative contents from a set of tabular documents (or web pages) of a web site. It proposes methods to automatically discover the intra-page redundancy and extract informative contents of a page.

Fig. 1: The Process of Info Discoverer

Info Discoverer, as shown in fig. 1, takes web page and partitions it into several content blocks called CB1, CB2, CBn based on HTML tag <TABLE> in a web page. Features in web page are all meaningful keyword except stop words (the, is, at, that). Statistics of occurrence of feature in each block is calculated, and then it calculates entropy value of each feature. Thentropy value of the feature appearing in web page is estimated by its weight distribution. The features of content blocks in a page are grouped and represented as a feature-document list with term frequency (TF) or weight as shown in fig. 2.

II. Comparative Study

A. Filtering Techniques for Content Extraction [4]
This system employs multiple extensible techniques for content extraction. To perform step of structural analysis and structural decomposition, the page is first passed through HTML parser creating a DOM tree. OpenXML is used as HTML parser, which handles correcting the HTML, therefore do not have to deal with error resiliency. The set of different filtering techniques are applied by extractor while navigating the DOM tree recursively. This is useful in removing and modifying specific nodes and finding the content in page. The model has two sets of filters, with different levels of granularity. The first set of filters simply ignores tags or specific attributes within tags i.e. images, links, scripts, styles, and many other elements can be removed from the web page.

The second set of filters provides a higher level of extraction, consists of advertisement remover, the link list remover, the empty table remover, and the removed link Retainer. The benefit of this algorithm is that it performs well on pages with large blocks of text such as news articles. As, implemented on proxy it spaws on a new thread to handle each new connection and also provides content extraction for group of people. The drawbacks are that it does not find the content but eliminates non-content. Also, the system needs to improve latency and scalability while serving many clients on proxy.

Fig. 2: Feature-Document Matrix for Calculating Entropy Values of Feature
The entropy of a content block, \( H(CB) \), is the average of all feature entropies in the block.

\[
H(CB_i) = \sum_{j=1}^{k} H(F_j)
\]

(1)

The content block can be divided into two categories based on \( H(CB) \) : redundant and informative. If \( H(CB) \) is close to 1 or higher than a defined threshold, the content block is redundant. Else if \( H(CB) \) is less than a defined threshold, the content block has informative content.

The benefits of this approach are that it solves the problem of intra-page redundancy. But the problem is the threshold which helps to determine informative and non-informative is not easy to determine. If the higher threshold is chosen, the higher recall rate is expected. However, the precision rate may become lower.

C. DOM-Structure Block Analysis [7]

The block analysis algorithm introduced can separate true content in a web document from hyperlinked-clutter such as text advertisements and long links of syndicated references to other resources. At first, it ensures that the retrieved document DOM is error-free. Then it starts from root node and recursively traverse its child nodes, noting information about Element, Parent and Structure elements. The algorithm determines all “structurally-important”, then establish which ones classify as potential roots of a sub-structure, take each one of these structures and their children in turn and analyse their content for hyperlinks. Then the ratio between non-empty anchor tags over the overall number of tags with a threshold of 0.5 is calculated along with the ratio of the length of hyperlinked text over the total length of text for each section with a threshold of 0.4. By this process the algorithm progressively identified areas of high hyperlink concentration within a web document and its attempt to separate them from the “core content”.

The benefit of this approach is that performed well in situations where pages contained one or more bodies of text along with hyperlink clutter. Also it’s able to correctly identify lists of hyperlinks that were constructed using tables (simple or nested), \(<DIV>\) tags, symbols such as the pipe (|) character or bullet-pointed lists. The drawback of this approach is that it’s less efficient when used in front pages of portals or other such web documents where an entire page is filled with short summarisations and their hyperlinks without giving prominence to any one body of content.

D. Statistical Model for Content Extraction [8]

In this approach the DOM tree is constructed for extracting the content from web page. Then the features like the quantity of text and the quantity of hypertext present at different nodes is analysed to determine the usefulness of each node in tree. The deviation and normalized deviation is calculated. The higher the deviation more is the information rendered by the node. Eq (2) for calculation of deviation

\[
Dev(i) = \frac{TextQuantity(i)}{MeanTextQuantity(t)}
\]

(2)

To normalize the result to a closed interval [-1, 1] use the Eq (3)

\[
NDev(i) = \frac{Dev(i)}{MaxDev(t)}
\]

(3)

Then to distinguish the informative nodes that have less quantity of information in DOM tree as for example the titles of the columns in tables, headings, calculate the link densities associated with the nodes of DOM tree along with their deviations. The link density for any DOM node \( i \) can be calculated as given in Equation (4)

\[
Density(i) = \frac{LinksQuantity(i)}{TextQuantity(i)}
\]

(4)

The benefit of this approach is that it has High accuracy gain. The drawback is that as formatting Characteristics are not considered, it fails to extract the content when quantity of content at any part of document is low.

E. Content Extraction from HTML Documents [1]

The introduction of Personal digital assistant (PDA) has given a need to access the information anywhere, at any time with consuming low bandwidth. This approach mainly deals with extracting the contents for PDA and other devices. It takes each web page decompose it, determine the relationship among content and summarize it. The steps for Content Extraction from HTML documents are as follows:

1. Segment the Web Pages Into Zones
   - Analyze the relationship of this zones:
   - Achieved by:
   - Proximity analysis
   - Content classification.

2. Steps in Extracting the Contents are:
   - Structural analysis
   - Decomposition
   - Contextual analysis
   - Summarization
   - Table of Content
   - Order of TOC

F. Content Extraction by Tag Ratios [9]

Content Extraction via Tag Ratios ( CETR) is a method in which HTML document’s tag ratios are evaluated to extract content from diverse webpages by using the. The tag ratios are calculated line-by-line basis and then cluster the resulting histogram into content and non-content areas. Three clustering approaches are introduced they are CERT, CERT-KM, CETR-TM. CERT-TM applies minimum cut-off or threshold (CERT-TM). CERT- KM partition using clustering approach based on similarities from the tag ratio alone. The Tag Ratio are the number of tags per line of HTML. Tag Ratios (TRs) are the basis by which CETR analyses a webpage in preparation for clustering. TRs are the ratios of the count of non-HTML-tag characters to the count of HTML-tags per line. If the count of HTML-tags on a particular line is 0 then the ratio is set to the length of the line.

In the TR algorithm D is the document which is analysed and TR is histogram created by applying the algorithm as shown below:

1. Input document D
2. Remove Script tags from D
3. Remove remark tags from D
4. Remove Style tags from D
5. For all \( i = 1 \) to \( D \) do
   - \( x = \) non tag characters in \( Di \)
   - \( y = \) tags in \( Di \)
   - if \( y = 0 \) then \( y = 1 \)
   - \( T(i) = x/y \)
After TR histogram, a Gaussian smoothing pass is made on histogram to calculate T it helps to retain the important contents in web page. The high tag ratio portion to be indicative of the webpage’s content.

Example Below is a brief snippet of a webpage newsarticle.

- `<div id="topnav">`<br>
- `<div id="storyPageContent2">`<br>
- `<div id="author">James Smith</div>`

This example was used to illustrate the tag ratio and content code blurring concepts.

The Tag Ratios for these six lines are computed as follows:
- Text=0, Tags=1, TR=0
- Text=0, Tags=1, TR=0
- Text=11, Tags=2, TR=5.5
- Text=37, Tags=0, TR=37

Threshold Method(CERT-TM): The threshold τ is determined which classifies the Tag Ratios into content and non-content sections. That is any TR value greater than or equal to τ should be labelled content, and conversely, any TR value less than τ should be labelled not content. The problem then becomes a task of finding the best value for τ.

Selecting Content via Clustering (CERT-KM): Alternatively, the k-means clustering method is to group content C and non-content N lines by using Tag ratios as the only similarity measure. Empirically, set k ← 3 number of clusters. The resulting k clusters S1, S2, ..., Sk are labelled by selecting the cluster which has its centroid closest to the origin Smin and assigning N ← Smin. C is assigned the remaining clusters. All HTML tags in content-lines C are stripped and output is created. This 1-dimensional k-means clustering method is hereafter referred to as CETR-KM.

The benefits of this approach are that it is a viable and robust content extraction algorithm. It performs well even on non-news bodies and across multiple languages. It achieves better content extraction performance than existing methods works well across varying web domains, languages and styles.

The drawbacks are that in some webpages wherein the HTML mark-up is written in a single line CETR would be forced to either return all text or no text. Even though CETR-TM does not perform the best overall, for practical purposes end users may consider its use when recall is a top priority. CETR does not perform well on portal home pages. In CETR the recall is high and precision is low. The webpages which do not have advertisements or menus, such as computer science professors’ homepages, do not achieve high extraction accuracy.

G. Content Extraction Using Content Code Blurring [3]

Content code blurring is a novel Content Extraction algorithm. It works well with invalid or badly formatted HTML documents. It is fast and robust in its extraction performance. It delivers very high extraction accuracy.

Works well with invalid or badly formatted HTML documents. It is based on considering the typical visual features of main and additional content i.e. highly formatted areas and areas that contain little and short texts. Aim good results on most documents. It is fast and robust in its extraction performance. It delivers very high extraction accuracy.

4. For each single element in CCV determine content to code ratio (CCR). It helps to find if it is surrounded by content or code.

5. If CCR=high then surrounded by text (main content).

Calculation of CCR: In CCV each entry is initialized with a value of 1 if content and with a value of 0 for code. To obtain the CCR calculate for each entry a weighted and local average of the values in vicinity with a fixed symmetric range. If all the elements in this vicinity started with a value of 1, the average will be 1; same for value 0. The whole process corresponds to constructing a one dimensional image from the atomic elements, in which each pixel represents a single element and is initially coloured white if it represents content and black if it represents code. The Gaussian blurring filter is applied to calculate the weights of neighbourhood so named as content code blurring. CCV is said to have main content if CCR values are above fixed threshold. This approach is specialized for wiki style pages.

III. Efficient Contentextraction Using Hybrid Technique

The paper proposes a hybrid model named as “Info Optics” (IO). This approach is called hybrid because it mainly operates on two models of content extraction one based on statistical features and other on Formatting characteristics. It operates on DOM tree representation of web page calculating the different statistical features associated with the different nodes of tree to measure their importance in providing the information. The different statistical values like text density and link density for each node are calculated. These values are normalized, so important content should be retained. The calculation is based on the fact that the nodes associated with the content have higher values for the quantity of the text and lower values for quantity of hypertext.

In order to achieve an optimal performance on different styled web pages quantities are normalized with respect to each page. Once the statistically useful nodes are identified, other nodes similar to useful nodes based on formatting characteristics and their position in the page are identified. All of the nodes classified as useful and nodes similar to useful nodes are considered to be the nodes containing real contents.

IV. Analysis of Content Extraction Techniques

The analysis of different content extraction techniques are shown below in table.

<table>
<thead>
<tr>
<th>Ref no.</th>
<th>Technique Used</th>
<th>Merits</th>
<th>Demerits</th>
<th>Type of Web Pages Handled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref[1]</td>
<td>Structured Analysis, Decomposition, summarization</td>
<td>Summarize web content on fly</td>
<td>Not organizing information in meaningful format</td>
<td>Can handle multi column layout</td>
</tr>
</tbody>
</table>
The Hybrid approach for content extraction is beneficial than previous techniques. It does not make prior assumptions on the web page structure. It is composed of two techniques statistical feature and formatting characteristic extraction, which help in efficient content extraction. Considering the formatting characteristics serves as a guide to understand the template of underlying web pages, this add and advantage that the model works well with different styled web pages. It also helps in identifying the less informative nodes.

V. Conclusion
This paper discusses different content extraction techniques, featuring their approach each with advantages and disadvantages. It also introduces a new hybrid method for content extraction and summarizes the benefits of its techniques over the older content extraction techniques.

References

G. Naveen Sundar, received the B.E degree in computer Science and Engineering from C.S.I Institute of Technology, Thovalai in 2002. He received the M.Tech degree from Karunya University, Coimbatore in 2006. He is currently working toward the PhD degree. He is working as an assistant Professor in Computer Science department of Karunya University. His main research interests are Association Rule Mining, Databases and Web Mining.

Sheba Gaikwad pursuing her M.Tech in Software Engineering from the Department of Computer Science in Karunya University received her B.E. in 2011 from North Maharashtra University.