Web Page Segmentation

Mahmood Ahmad, Sungyoung Lee
Department of Computer Engineering, Kyung Hee University, South Korea

Abstract

This paper describes an overview and research work related to web page segmentation. Over a period of time, various techniques have been used and proposed to extract meaningful information from web pages automatically. Due to voluminous amount of data this extraction demanded state of the art techniques that segment the web pages just like or close to humans. Motivation behind this is to facilitate applications that rely on the meaningful data acquired from multiple web pages. Information extraction, search engines, re-organized web display for small screen devices are few strong candidate areas where web page extraction has adequate potential and utility of usage

1. Introduction

The purpose of web page was envisioned as a unanimous platform of information about certain entity that exists anywhere in the world that can be accessed across the globe 24/7. With voluminous outburst of information, its usage and applications, has reshaped old concept of just accessing information about single entity when internet applications are realized as utility computing. Abundance of information that gets accrued over the web pages, user contributed contents, interaction with social media, tweets, blogs, weather and stock forecasting applications and email clients are main supporting pillars of this technology shift. A single user anywhere in this world can get a comprehensive information regarding any topic of interest over few clicks. Search engines are one of the most famous area of applications that have brought exact knowledge retrieval against user search query

Searching webpages is different from searching an organized record sets i.e, database. IN database record sets are organized into rows and columns and search query retrieve only the matching records. IN web pages, records are not organized like relational database as it is unstructured information. Secondly, closely matching ideas are also retrieved or suggested to user. These two aspects of information retrieval from unstructured data made web search applications to think beyond the conventional search mechanism. Likewise other applications that rely on abundant information that exist over the internet have to extract new but meaningful and related information from the web. Era of smart applications introduced smart devices like handheld devices, smartphones, navigation and palm devices. Provisioning to access internet through these devices have not only boosted up their usage but has demanded careful analysis while extracting the information from these web pages

A regular web page comprises on various sections like headers, side bars, footers, contact information panel, advertisements and main body contents. Although a large collection of web pages follow some certain template to design their web pages; however, still many of them are not compliant with W3C standard to publish a web page. Regarding a static web page, extracting information is relatively easy as compared to those web pages that are dynamic. Identifying the section of a particular web page that is susceptible to any change is one challenge and identifying that either that section has really been changed is another challenge. Usually not all, but few segments in a dynamic web page get changed over a period of time. It's not a processing cost to extract that updated portion of webpage, but making it cached on a web server is another challenge. Depending upon all these challenges and requirements web segmentation has become a hot research are due to dependency of various systems on the extracted information. Relying on human factor to extract and segment a web page is not only hectic but also impossible. Therefore various techniques have been proposed over a period of time for the extraction of web segments and useful information

2. Literature Review

We have included literature ranging between 1997 and 2009 for web page segmentation and the discussion has been done in the increasing order of years. Each work has been reviewed in terms of general overview, methodology and applications

2.1 Template based information mining

For mining un-structured information out of a web page, the very first work that we have analyzed [1] presents a methodology that exploit the structural template of a web page. Structural hypothesis has been proposed long ago to effectively utilize the web information

Since there is no formal specification to follow while publishing anything on the web, this idea never gets materialized. In this paper, the author claims that unstructured data inside a webpage is sufficiently structured for extracting meaningful contents. For this purpose the authors proposed tree structured templates. The information is extracted from a web page which is then mined through FAQ documents. According to paper, certain collection of web pages shares a similar logical structure. Depending upon the structure, content and format of the document (which are assumed to be the main ingredients of any document), information can be extracted. In problem formulation the web page(document) is represented in terms of five tuple approach that are, set of structural components, sequence of contents and formatting, partial order on structural components and a mapping function. An approximate tree matching algorithm has been developed to extract information from unstructured web pages. The document that

is under experiment is subject to test with a range of document templates and selects the one that matches best. After matching process, the meaningful information is extracted as defined by the tags. The aim of experiment to test either the proposed algorithm detects accurately QA section of web page for useful mining purpose on question and answers. After testing the collection of data set, only 65% of test data has been found accurately covered by the proposed algorithm. The rest of 35% documents which are not detected by the algorithm are due to multi-segment, unusual-format, ambiguous clarity of table of contents and only questions and answer pair.

2.2 Vision-based page segmentation

The work presented in [2], mimics the human like identification and classification for web pages. The proposed algorithm looks and divides the web page using visual and spatial cues. Therefore instead of relying on the DOM tree or HTML tags, using the page layout feature authors have proposed VIPS(Vision based page segmentation). Before building the visual page segmentation, the algorithm starts by extracting nodes using DOM tree and then identifies the separator (horizontal or vertical lines). At this point, the algorithm is now ready to perform semantic segmentation using the top down approach.

A particular web page under analysis or investigation of VIPS algorithm is identified through separators (horizontally). In each separated block, it is further investigated for any vertical separator. After identifying the separators, degree of coherence (DoC) is measured to check how coherent the contents within a block (boundary block with in the separators) are. Applying these method recursively within the newly identified blocks, the coarse granularity for page segmentation is achieved. After all blocks are processed, the final vision-based content structure for the web page is outputted. Every web page does not follow the W3C standard for authoring any web page, therefore relying only on HTML tags can falsify the output result, and therefore, for visual block representation its DoC value is set according to the intra visual difference. Later the authors define few cues which will help the proposed methodology. These cues are, Tag, color, text and size. These cues are then used to formulate a set of rules (13 rules in total) for the block extraction. To find and distinguish between different semantics that are there within a web page, separators (vertical and horizontal) are found helpful. These separators are assigned with different weights, for example, if separator has different color on neighboring blocks, then its value is set to high. The page is partitioned based on visual separators and structured as a hierarchy. This semantic hierarchy is consistent with human perception to some extent using the VIPS.

2.3 Automatic information extraction from semi-structured web pages by pattern discovery

To extract data from semi-structured web page and to get rid of existing wrapper methodology, the authors in [3] have proposed a pattern discovery algorithm and that too without training example. This component is called the extractor and named as IEPAD (Information Extraction based on Pattern Discovery). To start with, any web page is analyzed using IEPAD where its pattern is discovered through pattern discovered component. This pattern is then applied and tested on webpages which are similar to it. This pattern is nothing but a class of regular expression. While expressing the regular expression the authors have used predefined HTML tags, called tokens and the actual data inside these tokens as text strings. Text level tags are further divided into three categories including logical tags, physical tags, and special tags for marking up text in a text block. To identify repeated pattern within the web page PAT tree is constructed. Most information that is generated based on some predefined templates and is commonly aligned regularly and contiguously and to display patterns two measures, called ‘‘variance’’ and ‘‘density’’, are defined to evaluate whether a maximal repeat is a promising extraction pattern or not. To handle patterns with variance greater than the specified threshold, the occurrences of a pattern are carefully clustered to see if any partition of the pattern’s occurrences can form a regular block in which the pattern has a variance less than the threshold. Since PAT trees compute only ‘‘exact match’’ patterns, templates with exceptions cannot be discovered by PAT trees. Therefore, authors applied another technique called multiple string alignment to handle inexact or approximate matching. Since record patterns only tell where the records are located, there must be some way to designate attributes in a record therefore, function of the pattern viewer is to allow users to assign attribute names and information slots in a record. To extract finer contents, the idea of multiple string alignment and block division are employed. The retrieval rate for the testing set reaches 96.35% for one training page and 98.65% for two training

2.4 DOM-based Content Extraction of HTML Documents

Extraction of meaningful web contents for various application like handheld devices, speaking apps for visually impaired person is discussed by authors [4]. The hurdle to extract this information is posed by popups, unnecessary adds, and irrelevant images and links. To avoid this different techniques have been proposed like WPAR , Webwiper, and JunkBusters. These techniques either distort the original contents or orientation of original contents thus making it hard to read and understand. The authors have used DOM tree methodology to get rid of this un-necessary stuff which is not useful otherwise.

The authors select a web page that has to be extracted for its contents. This web page is then given to http://www.openxml.org (domain does not exist anymore) to construct the DOM tree. This tree is then processed with two steps. In first step images, links, scripts and styles are removed. In second step which is considered to be more complex is used to convert the rest of HTML into WML. Second step also removes advertisements, link list and empty tables. List of links through href and src are also investigated against advertisement server and blacklists. This black list is updated periodically on host side. After performing these steps, the remaining is the actual content that can be used for either information retrieval, applications for blind people or even for devices having small display area

2.5 Adapting web pages for small screen devices

Browsing applications are not limited with large display devices like computer monitors. In recent years, hand held devices and other mobile devices with smaller display size are also connected with the internet. To facilitate browsing on small sized screen is the goal of this paper. To transform a regular web page that is primarily designed for computer monitor or LCD screens, it has to be re-adjusted in such a way that can fit on smaller screens. Instead of creating and designing web pages from scratch, the authors [5] have designed a methodology that will transform an existing web page for mobile devices screen. Identification of various sections on a web page, DOM object and inside detail of header, footer and side columns of web page has been devised out in proposed methodology to achieve the goal. They have used 50 different websites with 200 different web pages to identify different sections of a web page. Possible location and size of various components on a web page was the information of interest to know. Nonlinear support vector machine with radial basis function The data has been collected from 50 different websites for 200 web pages in total. Their aim was to achieve possibility that the proposed algorithm can transform a web page for small screen devices. Besides this main goal, they also considered possibilities to deploy their model either on client, server or proxy side. These possible options were mainly due to compute power of mobile devices and network bandwidth usage.

2.6 Web data extraction based on partial tree alignment

Using two step approaches of visual segmentation and partial alignment, the authors have proposed an automatic mechanism [6] to extract data from web page. The web page is analyzed using HTML tags and DOM tree to identify the data inside it. To avoid errors, the HTML page is loaded into the browser and is given a visual analysis. To combine data which can be scattered across various tree, all sub trees are merged into one tree. This is done using the partial alignment method. To ensure high accuracy, only those data fields that can be aligned with certainty. Authors called this technique as DEPTA (Data Extraction based Partial Tree Alignment. The proposed algorithm works in three steps. Building HTML tag tree, Mining data regions and mining each data record in each data region. In already proposed idea of MDR algorithm, page visualization is added component in new work. Each element of HTML is identified by its four coordinates and then containment relationship between any HTML tag. The experiment is conducted with 72 web pages from 49 websites. The output of their experiment result is more than 98% both in Precision and Recall. Utility of the proposed model has great utility for applications that rely on data mining.

2.7 A densitometric analysis of web template content

Other than normal text that appears and covers a major portion of web page, contents that are driven by template may be or may not be useful for search engine. The author has given his idea for clear segregation between the actual content of web page and segmented contents. The author in [7] has used field of Quantitative Linguistics to corroborate his claim in classifying web materiel either as regular text content or allied information that appears through template. The author has collected his data and performed BlockFusion segmentation algorithm and then used the beta distribution as a fuzzy classifier between different classes of web page contents. The analysis is conducted on the representative Webspam UK-2007 dataset Adopted Methodology: BlockFusion segmentation algorithm and Beta Distribution Model. 56437 web pages have been put to test for the evaluation purpose. BlockFusion segmentation algorithm is used to check how close the methodology resembles with the manual classification. Using Beta distribution that author further divided his experiment into fuzzy classification. Aim of this fuzzy classification is to find inclination trend of any term either as actual text of web page or otherwise. Using degree of typicality that is ranged from [-1, +1], the terms are checked either as a member of class "C1" or "C2".

2.8 Extracting Article text from the web page with maximum subsequence segmentation

Instead of relying on time demanding manual efforts or compute intensive tasks for useful information extraction from web pages, the authors of this paper [8] have presented a semi supervised approach that is least both in computation and human engagement.

First they selected the data set to be evaluated on their proposed methodology. After selecting this data underwent to be tokenized using simple HTML tags. These token are then put for maximum sub sequence followed by identification for longest and smallest cell through naive Bayes classifier. They used popular news websites and used their web pages for both training and evaluation example Using HTML tags to identify various sections of web page, Maximum subsequence segmentation, Naive Bayes classifier. The data is classified into two sets i.e, training set and evaluation examples. 2000 training examples were collected from 12 news websites. 450 evaluation examples were collected by 45 web sites. In experiment section of the paper, authors have stated that they have used 24,000 labeled training examples. To identify and extract useful content of information out of web page in terms of Precious, Recall, and F1 score. In terms of precision and recall, mainly the results appeared to be more than 90%. Expecting this much result advocates the proposed methodology and its effectiveness. Relaxation of supervised learning and its shift to semi supervised learning is claimed as least human involvement for actual content identification. The candidate applications that can get benefit from proposed strategy are devices with small screen, RSS feeds, Information retrieval and link analysis.

2.9 Automatic fragment detection in dynamic web pages and its impact on caching

Requirement for updating static web page in web cache is far lesser than a dynamic page. In dynamic web page, not all contents are updated frequently; however, a complete refresh on it is required anyway. In this paper the authors have optimized web caching by updating only those fragments that are modified and changed and to achieve this identification of these fragments is first task in hand. To achieve this goal, the paper presents a methodology for efficient page fragment detection and web cache utilization.

To achieve the goal, the proposed technique first identifies the contents from within a web page that are dynamic and their occurrence in other pages. To detect either a fragment is shared by other pages or not, an indigenous algorithm is developed and named Shared Fragment Detection Algorithm. Further, the life time of any segment is also identified for better usage of cache. For caching any fragment, it is checked either its sharing in multiple pages or its lifetime of usage. These two identifications are main steps to fulfill the proposed idea methodology. Usage of Shingles tree instead of DOM tree is due to three reasons as described in the paper. First, it is more compact representation to construct the tree structure of any web page. Secondly, this method notices a slight change of contents within the nodes of web page, which is very useful to check which component or fragment is changing more frequently. Third one is the additional information tagged with each node of webpage which help for various comparisons. The advantage of shingle encoding over MD5 is its capacity to notice any change within a web page at more granular level. MD5 can change the entire representation of webpage after a slight change thus making it hard to track that which fragment was the actual reason. In their proposed methodology, html tags that are used for text formatting are not used.
The techniques used in their approach are; Augmented Fragment Tree (AF Tree), automated fragment tree with shingles and Shared Fragment Detection Algorithm (indigenous)
The experimental data has been taken from different versions of web pages from the web sites of BBC, IBM's portal for marketing, internet news, and Slashdot

2.10 Web page DOM node characterization and its application to page segmentation

The goal of this paper [10] is to extract unstructured data from a web page by ignoring its other segments like navigation bar, advertisement banner, headers etc.
To achieve the goal of this paper, the author has first used the DOM tree (also called as tag tree) to analyze the tree representation of a web page. The nodes in this tree are labeled as segments of web page in such a way that can be distinguished from other components of web page. For this purpose, the author has used content size and entropy to identify desirable segment of page. Size of the contents helps to identify either a node is web segment or otherwise. Here, the entropy mean, how much repetitive structure may appear in any node. The greater the repetitive structure (like bullets or hyperlinks) the greater the entropy of that node would be. Nodes with greater entropy are not considered as actual content of web page. Using these two parameters of content size and entropy, the author then deals with the node feature space. All these steps helped to conclude that there exist a correlation between the size and entropy of nodes. Based on this assumption the author presented his algorithm that starts with data cleaning, node feature computation and segmentation algorithm.
A collection of webpages from various domains of life, blogs, news, shopping, entertainment and sports
Content size and Entropy function (Author defined methodology), geometric progression relation
The input data is a collection of 400 webpages from Open Directory.
To find out the performance and accuracy of proposed algorithm for page segmentation the author used precision and recall. The aim is to test the algorithm in terms of accurately distinguishing the page segments from non-segmented content of web page.
The algorithm showed 90% precision at 80% recall. The chief drawbacks of the algorithm arises from the assumption that all page segments correspond to a single node in the DOM tree

Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2014-(H0301-14-1003)

3 References

[1] Jane Yung-jen Hsu and Wen-tau Yih Template-Based Information Mining from HTML Documents. " *AAAI/IAAI*. 1997

[2] Deng, Cai, et al. "VIPS: a vision-based page segmentation algorithm." *2010-11-21). http://www. cad. zju. edu. cn/home/dengcai/VIPS/VIPS. html* (2003).

[3] Chang, Chia-Hui, Chun-Nan Hsu, and Shao-Cheng Lui. "Automatic information extraction from semi-structured web pages by pattern discovery." *Decision Support Systems* 35.1 (2003): 129-147.

[4] Gupta, Suhit, et al. "DOM-based content extraction of HTML documents." *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.

[5] Chen, Yu, et al. "Adapting web pages for small-screen devices." *Internet Computing, IEEE* 9.1 (2005): 50-56.

[6] Zhai, Yanhong, and Bing Liu. "Web data extraction based on partial tree alignment." *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005.

[7] Kohlschütter, Christian. "A densitometric analysis of web template content." *Proceedings of the 18th international conference on World wide web*. ACM, 2009.

[8] Pasternack, Jeff, and Dan Roth. "Extracting article text from the web with maximum subsequence segmentation." *Proceedings of the 18th international conference on World wide web*. ACM, 2009.

[9] Ramaswamy, Lakshmish, Ling Liu, and Fred Douglis. "Automatic fragment detection in dynamic web pages and its impact on caching." *Knowledge and Data Engineering, IEEE Transactions on* 17.6 (2005): 859-874.

[10] Vineel, Gujjar. "Web page DOM node characterization and its application to page segmentation." *Internet Multimedia Services Architecture and Applications (IMSAA), 2009 IEEE International Conference on*. IEEE, 2009.