A NOVEL APPROACH FOR INFORMATION RETRIEVAL TECHNIQUE FOR WEB USING NLP

Rini John and Sharvari S. Govilkar

Department of Computer Engineering of PIIT Mumbai University, New Panvel, India

ABSTRACT

Webpages are loaded with vast and different kinds of information about the entities in the real-world. Information retrieval from the Web is of a greater significance today to get the accurate queried data within the desired time frame which is increasingly becoming difficult with each passing day. Need to develop a system to solve entity extraction problems from the web as compared to the traditional system. It's critical to have a clear understanding of natural language sentence processing in the web page along with the structure of the web page to get the correct information with speed. Here we have proposed approach for information retrieval technique for Web using NLP where techniques Hierarchical Conditional Random Fields (i.e. HCRF) and extended Semi-Markov Conditional Random Fields (i.e. Semi-CRF) along with Visual Page Segmentation is used to get the accurate results. Also parallel processing is used to achieve the results in desired time frame. It further improves the decision making between HCRF and Semi-CRF by using bidirectional approach rather than top-down approach. It enables better understanding of the content and page structure.

KEYWORDS

Information retrieval, NLP, Entity Extraction, Visual Page Segmentation (VIPS), Semi-CRF (Semi-Markov conditional random fields), HCRF (Hierarchical conditional random field) and Parallel processing.

1. INTRODUCTION

Natural Language Processing is an arena concerned with the interface between human natural languages and computer. The fields like Computer Science, Artificial Intelligence, and Computational Linguistics. Summarization, Sentiment analysis, Auto-categorization, search are many areas of the branches of Natural Language Processing. Entity extraction has gained importance in recent times due to information overload from tons of Webpages. This is happening because single entity information can be queried and the data about this entity can be obtained through thousands of Webpages. To improve people's browsing experience, it is extremely vital to understand the structure and semantics of Webpage. This paper explores the various developments in the field of information retrieval in Web using NLP.

It's an age of information overload where the user can get the required query through any medium. In this paper, we discuss the various techniques which can be combined and how it be improved through parallel processing. The introduction to NLP and the related work has been discussed in section II. In the section III, proposed approach of information retrieval technique for Web using NLP with parallel processing have been discussed and section IV concludes the paper.

DOI: 10.5121/ijnlc.2017.6104

2. Related Work

In this section, we have cited the relevant past literature that use the various information retrieval techniques for Web using NLP. Most of the researchers have combined techniques of this field to get the most effective results. We are in need of techniques which would focus more on the semantic portions in a web page. In the previous work in these areas, tag-tree is represented by tag structure which is primarily used to denote a web page. Instead of concentrating on the content structure more attention is given in the presentation structure.

Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma has proposed Vision-based Web Entity Extraction[2] using VIPS Algorithm (a Vision-based Page Segmentation Algorithm). Tag trees tend to focus on the presentation structure instead of the content structure which is the main issue with previous works as they often are not correct enough to differentiate in the web page's semantic portions. Also, designers have a different style to compose a web page which is intricate and varied. This paper proposes Vision-based page segmentation (VIPS Vision-based page segmentation (VIPS) approach for overcoming these problems. It does the page segmentation based on human perception and uses various page layout features like font-size, different colors used in the sections of a web page to build a vision tree intended for a page.

Jun Zhu, Zaiqing Ni, Ji-Rong We, Bo Zhang, Wei-Ying Ma has introduced a Hierarchical Conditional Random Field (HCRF)[3] model for understanding a page layout. To get efficient and accurate results on information retrieval of entities the importance of Page-layout understanding is an absolute necessity. With Vision-tree, nodes are the resultant output but assigning the labels becomes a task. It includes the long distance dependencies to achieve promising results.

Identification of entity is an important feature of information retrieval. To get required information for the specified query can only be obtained if the entities are well defined. William W. Cohen has introduced Semi-CRF [4] in which according to the assigned labels the text content inside the HTML element is segmented to identify the entities much better and accurate way. Also, we get the comprehensive portrayal about the entities as whole together. In the higher-order models, the computational cost is high which immensely reduced in Semi-CRF and it gives much of the outcomes of these models. It is an extension of CRFs. Instead of measuring the properties of individual elements it measures the property of segments.

3. Proposed Approach

We have proposed a approach which jointly combines three techniques Visual based page segmentation (VIPs), Hierarchical Conditional Random Fields (i.e. HCRF) and extended Semi-Markov Conditional Random Fields (i.e. Semi-CRF) along with parallel processing where the entire process in the background is run paralleled through concurrent processes to get efficient information in desired time frame. Now a day's information access through various mediums like desktop computers, laptops, mobile, tabs etc. are existing so to get the correct data within the efficient period is the need of the time.

Following is the framework we have proposed.

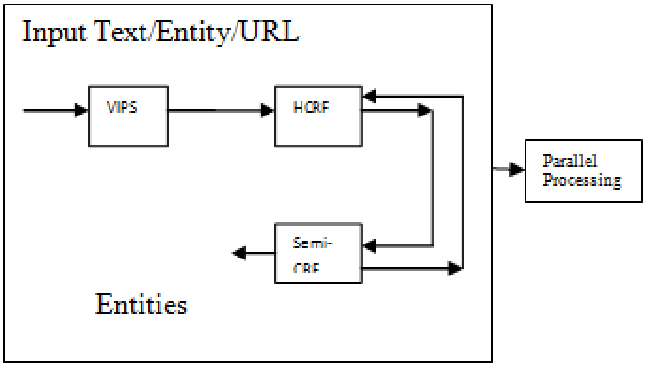


Fig. 1. A Basic model of Information retrieval for Web using NLP.

As shown in the above Fig. 1, the input is given as text file or Entity or a URL is given as the input to VIPS where Vision-tree is created based on the web page given where different rules are applied like the font, layout structure, separators which are explained in detailed below. Thus, this is given as input to HCRF where assignment and identification of each of the text context in the content structure of the node of HTML element are done and later Semi-CRF segments in the text context is further segmented to get accurate and better results and this is a bidirectional integration as seen in the above framework to process the data in an iterative manner. Also, all this process is done through parallel processing where the instructions are divided among various processors to get the information in the desired time frame.

In the end, the output is expected to be entities extracted from the particular entities given. After getting the entities, these entities can be further searched through the search engine for additional information.

For the further understanding of the approach following example can be taken into consideration.
Input: The input can be Entity or a Customized Text or the URL from which you want the entities to be extracted.

Input:

Enter Entity Name:

Enter Customized Text:

URL:

Output: Based on the Input, the entities are extracted and the type of entity whether it is a person, organization or location is identified and related information of each of the entity can further be displayed.

Output:

Entity Name:

Type:

Input: In the following example Obama is the entity to search so in the output Obama as a person is categorized.

For the customized text example taken is Obama is the president of America, the output is
Person ➔ Obama, Country ➔ America

Also similarly, below the URL is given in the input. So the output will be the extracted entities and type identified for each entity extracted.

Enter Entity Name: Obama

Enter Customized Text: Obama is the president of America

URL:http://en.wikipedia.org/wiki/LofNLPtools #Natural_language_processing_tool.

Type: Athelete / Person/Country

Information: Related URL

The proposed approach for information retrieval for Web using NLP consists of following phases:

- Vision-based Page Segmentation
- HCRF (Hierarchical Conditional Random Field)
- Semi-CRF (Semi-Conditional Random Field)
- Vision-based Page Segmentation (VIPS) phase[2]

The process for the construction of the vision tree from the content structure of the web page is given below. Information retrieval can take advantage from this page structure as VIPS uses a tag-tree free method to get the vision tree.

The VIPS Algorithm

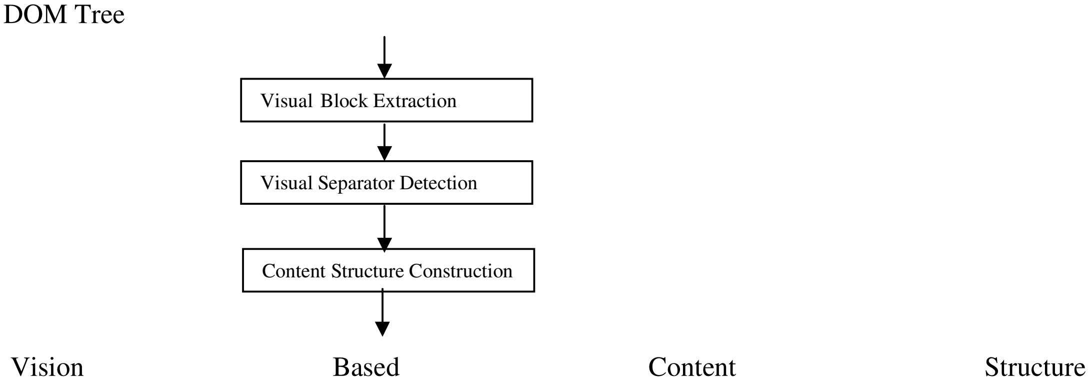

DOM Tree
Visual Block Extraction
Visual Separator Detection
Content Structure Construction
Vision Based Content Structure

Fig. 2. Visual Page Segmentation Process

In VIPS, there are three modules as shown in Fig. 2.

i. Visual Block Extraction: From the current subpage, the goal is to find the appropriate visual blocks. As per the intra-visual difference of every extracted node, the DoC value is fixed. Until all the proper nodes are found, this process is reiterated to get the visual blocks of the present sub-page.

Following is the Algorithm for Visual Block Extraction:

```
Algorithm PartitionDomTree (root, levelno)

{IF (Partitionable(root, levelno) == TRUE)

{For each node child of root {

PartitionDomTree (childnode, levelno); }}ELSE

{Create block in the pool by placing

 the sub- tree (root)}}

Algorithm Partitionable (root, levelno)

{IF (Top block is the root) {

RETURN TRUE;

} ELSE { Various rules are run for the following
HTML tags
FORM,UL,TD,P,TR,TBODY,TABLE;}}
```

The input is the DOM tree which consists of the visual blocks of the original web page. In VIPS, some the nodes can be further partitioned based on some huge nodes like <P> and <TABLE> as we get more focused visual blocks. Then these child nodes are created based on Heuristic rules like shape, size or color of the node.

For example tags <B>, <STRONG>, <FONT>, <I> etc. become the perfect candidates for further division of node. If one is a bold style and another line is of italics font style us as the user can perceive it's a different section of the same page. Thus, a division happens based on such rules. Another case can be considered of font size. The font size of the header would be different than the font size of the paragraph. Thus, these the rules applied in the VIPS algorithm to get the Vision tree.

Another case can be considered of font size. The font size of the header would be different than the font size of the paragraph. Thus these the rules applied in the VIPS algorithm to get the Vision tree. Following is the actual snippet of the code for the above logic used in the project.

```
private void InsertDOMNodes(IHTMLDOMNode parentnode,TreeNode tree_node)
                {
                        if(parentnode.hasChildNodes())
                        {
                IHTMLDOMChildrenCollection allchild =
        (IHTMLDOMChildrenCollection)parentnode.childNodes;
                                int length = allchild.length;

                                for(int i=0;i<length;i++)
                                {
                                        nodenum++;


                                        IHTMLDOMNode child_node =
(IHTMLDOMNode)allchild.item(i);
                                        TreeNode tempnode =
tree_node.Nodes.Add(child_node.nodeName +"_"+nodenum);
                                        InsertDOMNodes(child_node,tempnode);
            // holePage = tempnode.Text;
                                }
                        }
                        return;
                }
```

ii. Separator Detection: Visual separator detection algorithm is run to identify the separator between the block and relation among them; this is done for every block in the pool until we get the separated block.

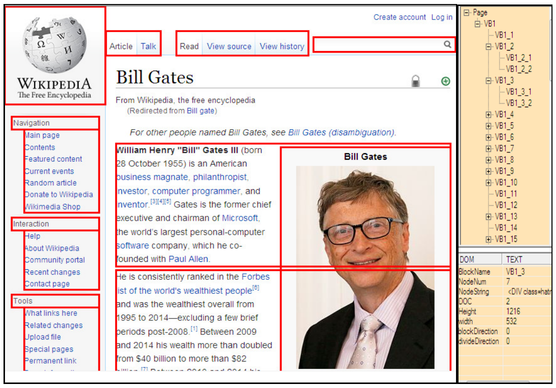
Separator detection process is needed for further block detection which is based on user perception as a user can perceive semantic division between different areas of the page based on the horizontal and vertical lines as they visually cross each other. In Fig. 3, example we can easily identify various sections of the page based on this concept. Image, paragraphs etc. are divided as they are vertical and horizontal lines are crossing each other. Thus, a separator detection algorithm is run based on this concept where weights to the separator are given based various parameters like the distance between the widths of the separator.

The visual separator detection algorithm is described as follows:

1) Initialize the separator list.

2) For every block in the pool, the relation of the block with each separator is evaluated
If the block is contained in the separator, split the separator;
If the block crosses with the separator, update the separator's parameters;
If the block covers the separator, remove the separator.

3) Remove the four separators that stand at the border of the pool.



Bill Gates

From Wikipedia, the free encyclopedia
(Redirected from Bill gate)

For other people named Bill Gates, see Bill Gates (disambiguation).

William Henry "Bill" Gates III (born 28 October 1955) is an American business magnate, philanthropist, investor, computer programmer, and inventor.[3][4][5] Gates is the former chief executive and chairman of Microsoft, the world's largest personal-computer software company, which he co-founded with Paul Allen.

He is consistently ranked in the Forbes list of the world's wealthiest people[6] and was the wealthiest overall from 1995 to 2014—excluding a few brief periods post-2008.[1] Between 2009 and 2014 his wealth more than doubled from $40 billion to more than $82 billion.[7] Between 2010 and 2014 his

DOM	TEXT
BlockName	VB1_3
NodeNum	7
NodeString	<DIV class=hatr
DOC	2
Height	1216
width	532
blockDirection	0
divideDirection	0

Fig. 3. Sample page of Separator detection and Vision tree

ii. Content Structure Construction: Content structures can be constructed when the weights are established and detected of the separators. Separators of the lowest weight and the blocks are criteria to form new blocks thus initiating the construction process. In the end, a vision tree with the various root, nodes, and corresponding vision tree is constructed as shown in Fig. 3.

1. Hierarchical Conditional Random Field (HCRF)[3]

In Hierarchical Condition Random Field (HCRF) helps in labeling that is identifying and assigning labels to the HTML elements. The following example can be considered for better understanding if city state and state is given in an address block

RUBY-503, BHOOMI BLDG,
SECTOR 35
MUMBAI MAHARASHTRA (MH) – 410209

Then using HCRF the following outcomes for the last line would be

CITY_STATE_ZIP-CODE
➔ MUMBAI_MAHARASHTRA (MH)-410209

The output of the HCRF model is the graph of the web page where junction tree algorithm is used to understand graph labels of vertices. The nodes of the vision tree are the vertices of the graph.

2. Semi-CRF [4]

Semi-CRF is used for segmentation of an input sequence and assigning labels to these segments. The following example can be considered to understand more clearly. An address line is given where the city, state and the zip code is given.

MUMBAI MAHARASHTRA (MH) – 410209

Then using Semi-CRF the following outcomes

CITY_STATE_ZIP-CODE

➔CITY → MUMBAI
➔STATE → MAHARASHTRA
➔ZIP-CODE →410209

It is an addition to the linear chain CRF in which iterative process of labeling is done for segments.

3. Information retrieval technique for Web using NLP with parallel processing:

As shown in Fig. 1, we have proposed a system for Information retrieval technique for Web using NLP with parallel processing where the above three techniques Semi-CRF, HCRF, and VIPS have been incorporated to get the extracted entities. The previous works have shown tag-tree dependent approach where more dependency is on the presentation rather on the textual content structure of the web page. Hence proper results can't be achieved as depending on the presentation structure can be risky because different designers have various styles of designing the web page. Thus, VIPS has overcome this issue and getting a vision tree based on the semantic portions of the tree.

Then further this vision tree is given to HCRF for labelling of the HTML elements in the child nodes of the tree as seen in the above examples. And then we are using Semi-CRF for segmentation of the text content to get finer and accurate results which can help to get the search results of the user query to the point. In previous works the top-down approach of HCRF to Semi-CRF have been implemented, the drawback of such approach is the results of the Semi-CRF cannot be passed to HCRF as it reduces the possible searching space. Hence bidirectional integration has been introduced to overcome this problem.

All this outcomes will be processed through Parallel Programming in the .NET Framework which enables to write efficient, scalable code which is divided among various concurrent processors to achieve the results quickly and accurately.

Following is the snippet code where we have used the parallel programming in which When you create a task, you give it a user delegate that encapsulates the code that the task will execute. This has been helpful as we can see the results with much less execution time as compared to not using Task Parallel Library.

```
using System;
using System.Threading.Tasks;

        // parallel programme
        Task t = Task.Factory.StartNew(() =>
        {
           NameEntityRecognization(holePage); });
```

3. CONCLUSIONS

In this paper, we have introduced information retrieval technique for the Web using NLP and parallel processing is used for efficient information extraction and retrieval. It's important to understand the page structure and layout plus the natural language to get the correct outcomes. Here in this framework, HCRF and Semi-CRF model are jointly implemented in the bidirectional way to get the better output through iteratively optimized technique. The results would be achieved in efficient time frame with the use of parallelism.

ACKNOWLEDGEMENTS

I would take this opportunity to express my profound gratitude to Prof. Sharvari Govilkar who has always been supportive and guided me throughout this project beginning with the idea of this topic. Her perspective has made this project better to understand and guide it to the right direction. I remain immensely thankful and grateful to work under her which has been a knowledgeable experience for me.

REFERENCES

[1] Rini John, Sharvari S. Govilkar., (2016) "Survey of Information Retrieval Techniques for Web using NLP", VOL.135, NO. 8.

[2] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "VIPS: A vision-based page segmentation algorithm", Microsoft Tech. Rep., MSR-TR-2003-79, 2003.

[3] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma, "Simultaneous record detection and attribute labeling in web data extraction", in Proc. Int. Conf. Knowl. Disc. Data Mining, 2006, pp. 494–503.

[4] S. Sarawagi and W. W. Cohen, "Semi-Markov conditional random fields for information extraction", in Proc. Conf. Neural Inf. Process. Syst., 2004, pp. 1185–1192.

[5] C. Yang, Y. Cao, Z. Nie, J. Zhou, and J.-R. Wen, "Closing the loop in webpage Understanding", in Proc. 17th ACM Conf. Inf. Knowl. Manage., 2008, pp. 1397–1398