# INFORMATION RETRIEVAL TECHNIQUE FOR WEB USING NLP

Rini John and Sharvari Govilkar

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

***ABSTRACT***

*Information retrieval is becoming an intricate part of every domain. Be it in acquiring data from various sources to form a single unit or to present the data in such a way that anyone can extract useful information and hence used in data analysis, data mining etc. This arena has gained much importance in the recent years because as of today we are exploded with various kind of information from the real-world. The growing importance of research data and retrieving the intelligent data are the main focus for any business today. So coming years this is a field where major work need to be done. We have focused here to implement a system for information retrieval from the webpages using Natural Language Processing (NLP) and have shown to getting better results than the existing system. Webpages is a home to huge amount of information from various entities in the real-world. Here we have designed a system 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

NLP technologies involve designing of algorithms and implementing them in a way that resembles to human processing for the similar task. It considers the way of interaction among the people and also machines processing to learn various rules and later apply those rules for the same task and to create intelligent data. The present formal structural models and how to incorporate them in the existing NLP technologies is the question to be considered in the future. Here we explore the various developments in the field of information retrieval in Web using NLP. The importance of web information has increased to number of folds due to advancement in access medium. So integrating and extracting various entity information has gained significance as of today.

Information retrieval is a process of getting required information in a desired set of time. This process depends on various factors to fetch accurate data in efficient time. The factors can be the system used for retrieving the data, how the data is stored and schema design etc. In this age of information blast we need efficient and new improved techniques to gather intelligent data in addition to the related data. Here we are exploring how to combine NLP techniques for getting better understanding from user's point of view and processing the same way as human do. Hence clear knowledge of semantic part of the data becomes critical to get a quantifiable improvement.

DOI: 10.5121/ijnlc.2017.6501 

## 2. RELATED WORK

In this section, we have cited the relevant past literature that use the various information retrieval techniques for Web using NLP [1]. 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.

Vision based Page Segmentation Algorithm is proposed by the authors Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Human perception is a critical aspect to understand user information and get better accurate results accordingly. VIPS does the page segmentation based on visual cues as per user perception. It is a Tag tree independent approach,

For acquiring better knowledge with respect to Page Layout of a web page, the author Jun Zhu along with co-authors have introduced Hierarchical Conditional Random Field (HCRF) [4]. 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.

Relevant entity identification plays a key part in information retrieval process. There are uncountable number of queries which are processed each day so get the accurate data for the given query which in turn has become utmost of importance. The author William Cohen introduced Semi-CRF [5] where once the entities are assigned labels to identify it more accurately and efficiently. This is an extension of CRFs. Here the measurement of property of segment is done.

To integrate the common desktop applications such as word processors, email clients , Web browsers),web information systems(wikis,portals),mobile applications and Natural language processing (NLP) techniques, the Semantic Assistants project[6] focuses directly bring it to the end users. To enable this integration, a service-oriented architecture has been developed that allows integrating clients with NLP services implemented in the GATE framework.

Paolo Nesi, Gianni Pantaleo and Marco Tenti [8] have presented Geographic information Extraction from Unstructured text data and web documents. For the organizations which are working on such domains extracting the correct geographical coordinates and addresses is highly expected of them.NLP techniques jointly combined with external knowledge (in the form of gazetters) have taken into consideration in this work.

The authors Suma Adindla and Udo Kruschwitz [9] have combined the best two worlds i.e. NLP and IR for intranet search. The authors has proposed a system that directs the user in the search process by getting the data collection and mining the pieces of knowledge from the documents which is automatically getting populated in database.

Zhong Liu and Ying Wang [10] have developed a novel method of Chinese Web Information Extraction and Applications for retrieving semantically rich information from the unstructured and semi-structured Chinese web pages, Zhong and Ying have presented a work flow of their IE system. To extract pattern and built knowledge repository the approach used are Knowledge engineering approach and automatic training.

Ruiqiang Guo and Fuji Ren [11] have analysed in this paper the link between NLP and Semantic Web. They have explained how NLP benefit Semantic Web to implement information retrieval.

B.Aysha Banu and Dr.M.Chitra [12] have proposed a novel ensemble vision based deep web data extraction technique for Web Mining Applications. Vision based approach is used to retrieve

information from the web pages. It assists the user instinctively to partition the webpage into number of semantic parts where the visual and spatial features are vital to this process. It emphasis on the primary visual features of a web page.

I.Vijayalakshmi and Sobha Lalitha [13] have provided a search facility for documents containing text which are in mobile or web and how to retrieve them is presented here. For the implementation they have considered English and Tamil documents for information retrieval.

## 3. PROPOSED APPROACH

We have proposed an new approach [2] 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 efficient period is the need of the time. Following is the framework we have proposed.

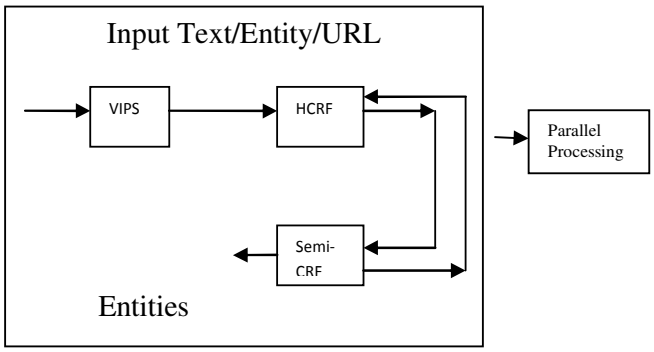



**Fig. 1.** A Basic model of Information retrieval for Web using NLP.

As shown in the above figure 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 input is given as the input to HCRF where assignment and identification of each of the text context in the content structure of the node of Html element is 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 a 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 further understanding of the approach following example can be taken into consideration. The input can be Entity or a Customized Text or the URL from which you want the entities to be extracted.

The proposed approach for WebNLP consists of following phases:

- Vision-based Page Segmentation
- HCRF (Hierarchical Conditional Random Field)
- Semi-CRF (Semi-Conditional Random Field)

**1) Vision-based Page Segmentation (VIPS) phase [3]**

The process for the construction of the vision tree [3] from the content structure of the web page is given below. Information retrieval can take advantage from this page structure as VIPS uses tag-tree free method to get the vision tree.

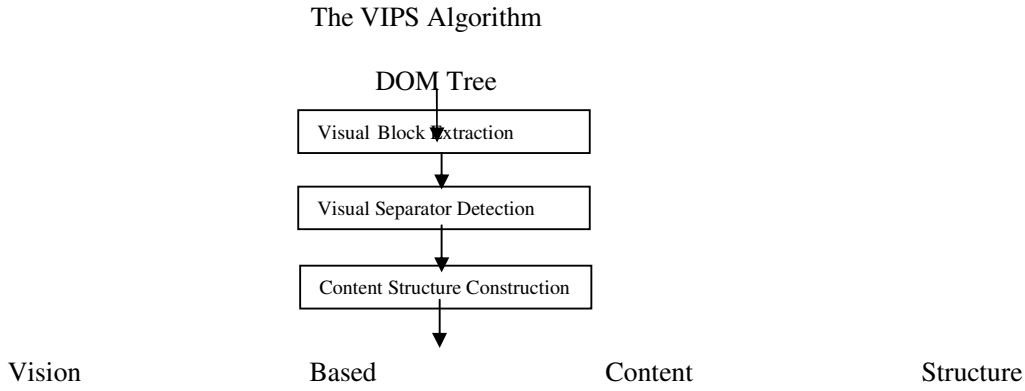



**Fig. 2.** Visual Page Segmentation Process

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

i. Visual Block Extraction: Here the aim is to find the visual blocks based on the semantic part of the web 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 colour 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 other line is of italics font style we as 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.

ii. Separator Detection: This 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.

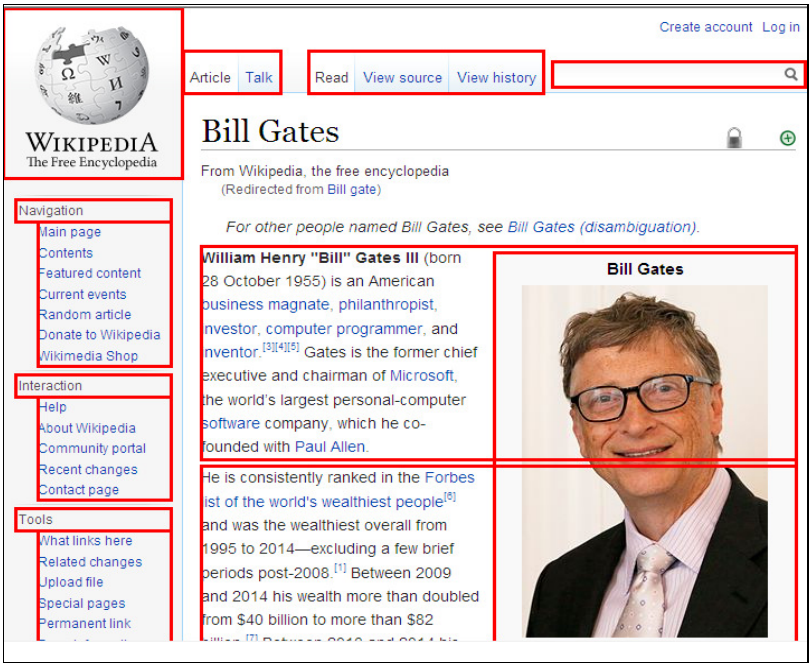
The visual separator detection algorithm is described as follows:

(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.



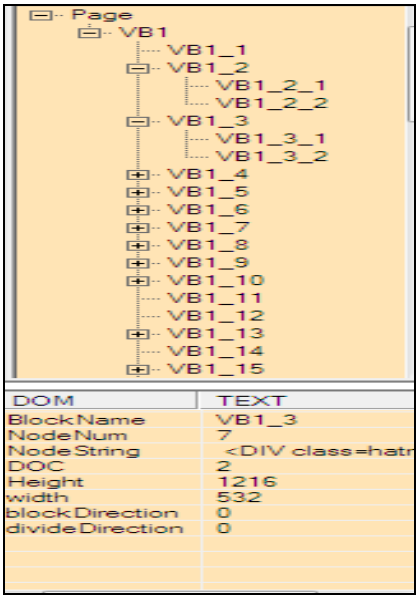



**Fig. 3.** Sample page of Separator detection and Vision tree

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.

**2) Hierarchical Conditional Random Field (HCRF) [4]**

In Hierarchical Condition Random Field (HCRF) helps in labelling 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.

3) Semi-CRF [5]

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

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

**1. 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[18]. 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.

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.

Below is the working the working model implemented.

Enter the URL or entity needed to be searched. Either a URL or entity can be entered, when the entity or URL is entered it will go to Wikipedia to search for the entity information. When clicking into the search button DOM tree is getting created. Here the entity below is Bill Gates.

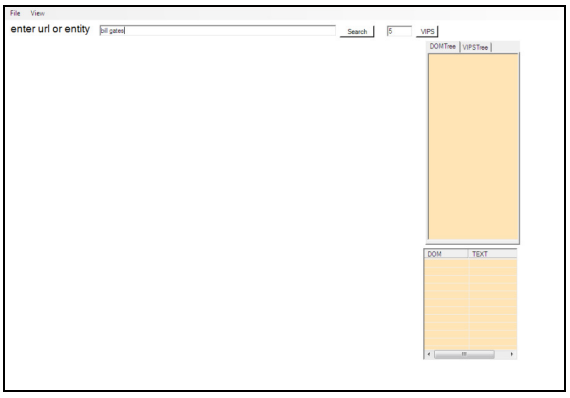



**Fig. 4.** GUI for input URL or Entity

Hence for the entity search “bill gates”, the system will go to Wikipedia and search for the information related to the particular entity and the segmentation process is getting started as shown below. Before the segmentation process starts the Vision tree is created with help of Visual page segmentation process where the page is segmented into semantic blocks which seen below output of the GUI.

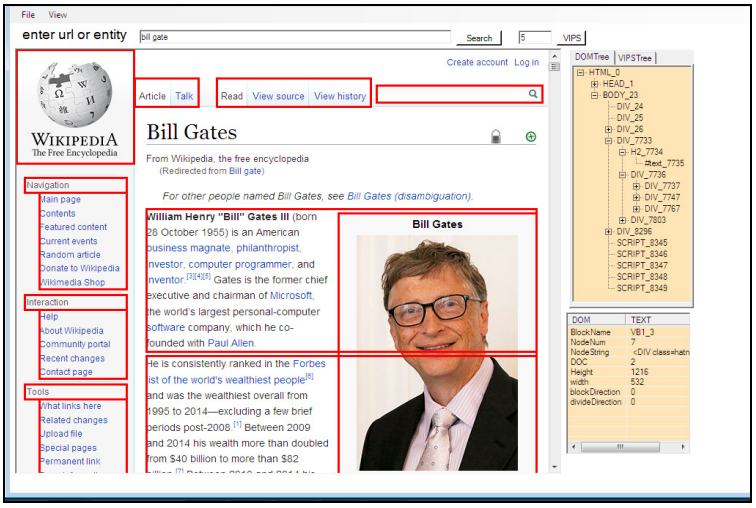



**Fig. 5.** GUI of the input segmented page

Also the corresponding VIPs tree is created. To view the entities in the corresponding webpage, click on the view entity on the main page. Here the nodes are created by applying various rules as far as possible to human perception such as the font style, the size, table, paragraphs.

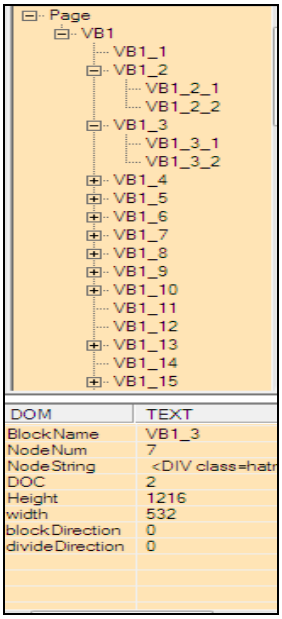



**Fig. 6.** GUI of the Vision Tree

So the following entities of the corresponding web page generated which as shown below. Also further information regarding the entity is required then further click on the entity so the desired

information is generated through the search engine. These entities are extracted from the semantic block structure of the vision tree. Through the technique of NER we are able to incorporate the logic of Semi-CRF and HCRF through which the entities are labelled and segmented to get the extracted entities.

| PERSON | ORGANIZATION | LOCATION |
|---|---|---|
| Abiogenesis | A&E Television Networks | Albuquerque |
| Alex | Apple | America |
| Alice Walton | ASR | Basa Jawa |
| Allen | Atari | Beijing |
| Amancio Ortega Gaona | Atari Program Exchange | Bellevue |
| Amanda | Austin Statesman | Breukelen |
| Amy Hood | Aztec Eagle | British |
| Andrew Carnegie | Bahasa Melayu | Cambridge |
| Anthony Michael Hall | Bank | China |
| Ballmer | BBC | Dallas |
| Barack Obama | BBC News | Fortran |
| Ben Bernanke | Berkshire | Iceland |
| Berkowitz | Berkshire HathawayReligionRoman Catholicism | India |
| BillAlternative | Berkshire Hathaway | Japan |
| Bill | Bill & Melinda Gates Foundation | Jordan |
| Bill Gates Gates | Bloomberg | King County |
| Bill Gates | Bloomberg Billionaires Index | Kirkland |
| Bina Abraham | Bloomberg Businessweek | Kristianne |
| Blair | Bloomberg NewsWorks | Lake Washington |
| Blakely | BNF | Langdell |
| Boies | Bower Award for Business Leadership | London |
| Bolger | British Computer Society | Manes |
| Bono | BusinessWeek | mankind |
| Buffett | Cambridge University | Mexico |
| Bunnell | Cascade InvestmentChair of Corbis | New Mexico |
| Carlos Slim | Cascade Investment LLC | New York |
| Chapman | CCC | Nigeria |
| Charles Koch | Church Hill Club | San Diego |
| Charles Sykes | CNET | Seattle |
| Christos | CNN | Silicon Valley |

**Fig. 7.** GUI of the Extracted Entities

These entities are searched on the basis of the attribute of Person, Location and Organization. Further search on particular entity can be done from list of entities extracted. For the above entities, the search of entity "Apple" will incur following results. The result of this captured through the Google search engine.

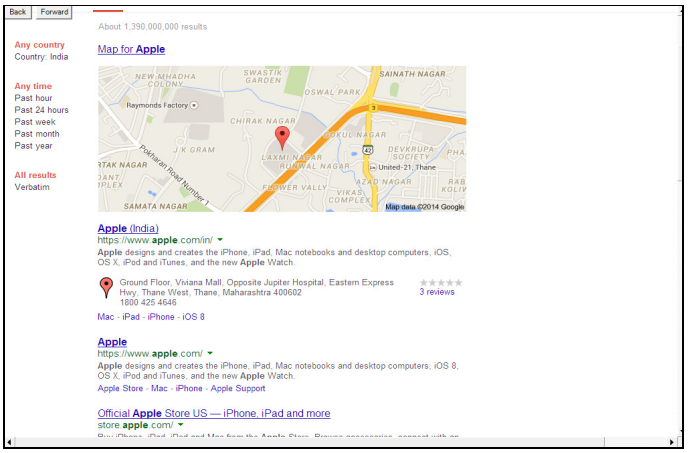



**Fig. 8.** GUI for the particular Entity search

Another option given is to extract entities through a file i.e. word file or text file. Below is the GUI for the uploading a text file. Here for the extraction of the entities the system directly process through NER, extracting the relevant Person, Organization and Location. As here we can see Vision tree is not created as there no design where division through human perception can be done. It's a text file where simply an entities are getting extracted.

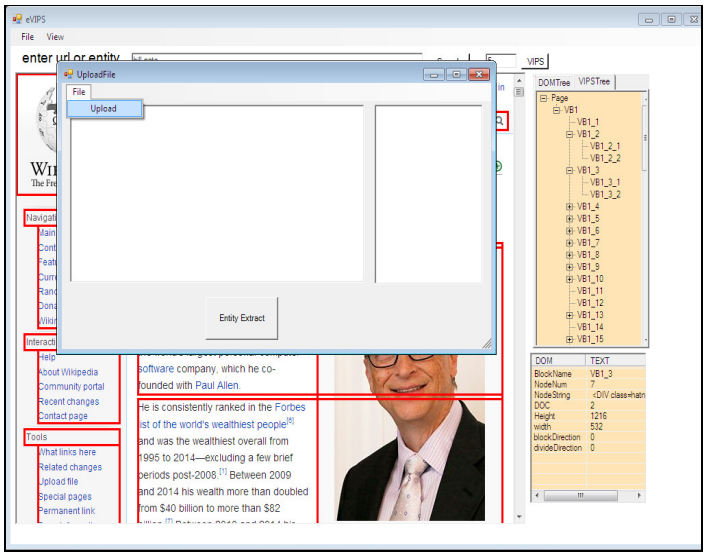



**Fig. 9**. GUI for the upload of text file

The extracted entities can be seen in the right side of the window of the below screen. Hence the output of the described attributes is getting generated through entity extraction.

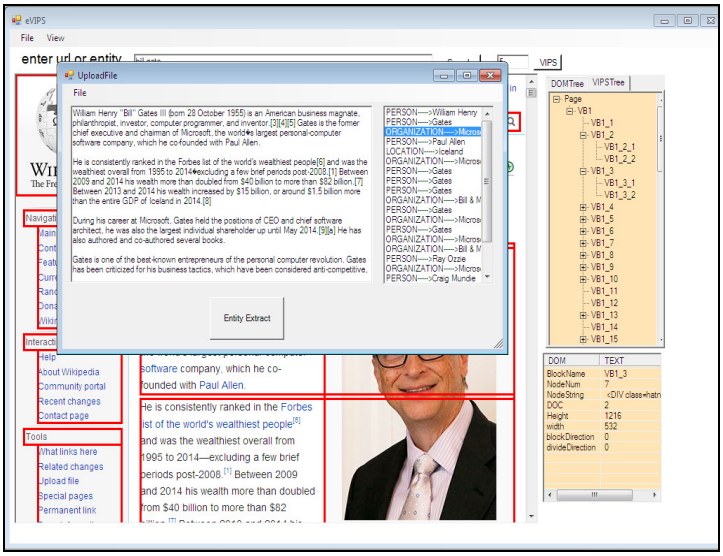



**Fig. 10**. GUI of the Extracted Entities

## 4. PERFORMANCE STUDIES

For accurately evaluating the system, there is a need for the measures of the actual performance of the system. In any Entity extracting system it is very much important to retrieve explicit and precise answer to which will satisfy the user. There are different performance measures available to evaluate question answering systems.

Evaluation is highly important for designing, developing and maintaining effective information retrieval or search systems as it allows the measurement of how successfully an information retrieval system meets its goal of helping users fulfil their information needs. But what does it mean to be successful? It might refer to whether an information retrieval system
retrieves relevant (compared with non-relevant) documents; how quickly results are returned;

**Table I**. Evaluation of System based on input as text

| Input Document | Total no. of words in the document | No. of actual relevant words(TP) | No. of actual irrelevant words(TN) | No. of irrelevant words which turn out to be positive (FP) | No. of relevant words which turn out to be negative(FN) | Recall (%) | Precision (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 112 | 18 | 93 | 3 | 1 | 95 | 86 | 97 |
| 2 | 236 | 35 | 196 | 11 | 5 | 88 | 76 | 94 |
| 3 | 282 | 68 | 212 | 14 | 2 | 97 | 83 | 95 |
| 4 | 155 | 20 | 132 | 5 | 3 | 87 | 80 | 95 |
| 5 | 294 | 93 | 194 | 17 | 7 | 93 | 85 | 92 |
| 6 | 160 | 80 | 62 | 10 | 18 | 82 | 89 | 84 |
| 7 | 89 | 30 | 54 | 3 | 5 | 86 | 91 | 91 |
| 8 | 65 | 26 | 37 | 4 | 2 | 93 | 87 | 91 |
| 9 | 120 | 53 | 63 | 6 | 4 | 93 | 90 | 92 |
| 10 | 173 | 89 | 79 | 11 | 5 | 95 | 89 | 91 |
| 11 | 52 | 18 | 31 | 3 | 3 | 86 | 86 | 89 |
| 12 | 116 | 57 | 52 | 7 | 7 | 89 | 89 | 89 |
| 13 | 69 | 29 | 35 | 4 | 5 | 85 | 88 | 88 |
| 14 | 41 | 18 | 18 | 2 | 5 | 78 | 90 | 84 |
| 15 | 83 | 42 | 36 | 6 | 5 | 89 | 88 | 88 |
| 16 | 91 | 33 | 56 | 5 | 2 | 94 | 87 | 93 |
| 17 | 58 | 18 | 32 | 2 | 8 | 69 | 90 | 83 |
| 18 | 64 | 34 | 27 | 5 | 3 | 92 | 87 | 88 |
| 19 | 89 | 36 | 49 | 3 | 4 | 90 | 92 | 92 |
| 20 | 126 | 79 | 37 | 12 | 10 | 89 | 87 | 84 |

how well the system supports users' interactions; whether users are satisfied with the results; how easily users can use the system; whether the system helps users carry out their tasks and fulfil their information needs; whether the system impacts on the wider environment; how reliable the system is etc. The parameters Precision, Recall and Accuracy are used to evaluate the system. These parameters are explained in detail below.

Precision and recall can be easily defined using a contingency table where TP indicates number of true positives; FP indicates number of false positives, TN indicates number of true negatives and FN indicates number of false negatives.

True positives are answers stated as relevant by both the human and the software. False positives are answers returned by the software, but were reckoned irrelevant to the question by the human. False negatives are relevant answers which are not found by the system. True negatives are the answers which are not returned by the system and are human also considered it irrelevant. Accuracy, precision and recall defined in terms of TP, TN, FP and FN is as follow:

Recall = TP / (TP+FN)
Precision = TP / (TP+FP)
Accuracy = (TP+TN)/ (TP+FP+TN+FN)

Following is the evaluation of the extracted relevant input entities from the text documents with the help of parameters Recall, Precision and Accuracy along with percentage of these parameters compared right below:

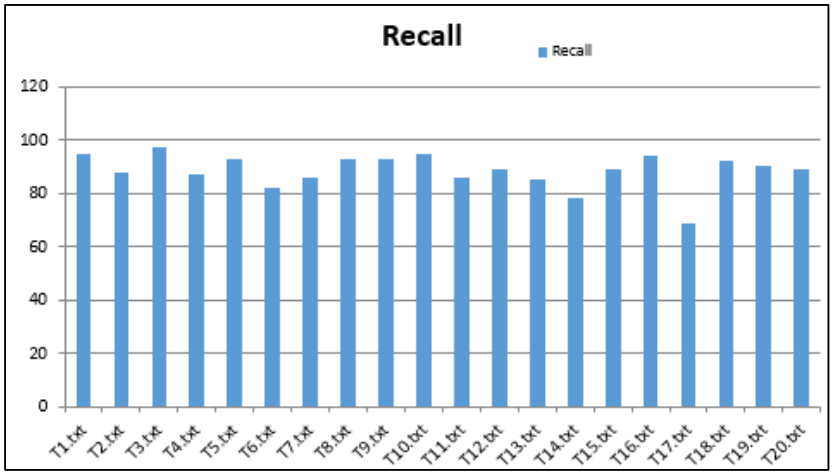



**Fig. 11**. Percentage of Recall Achieved

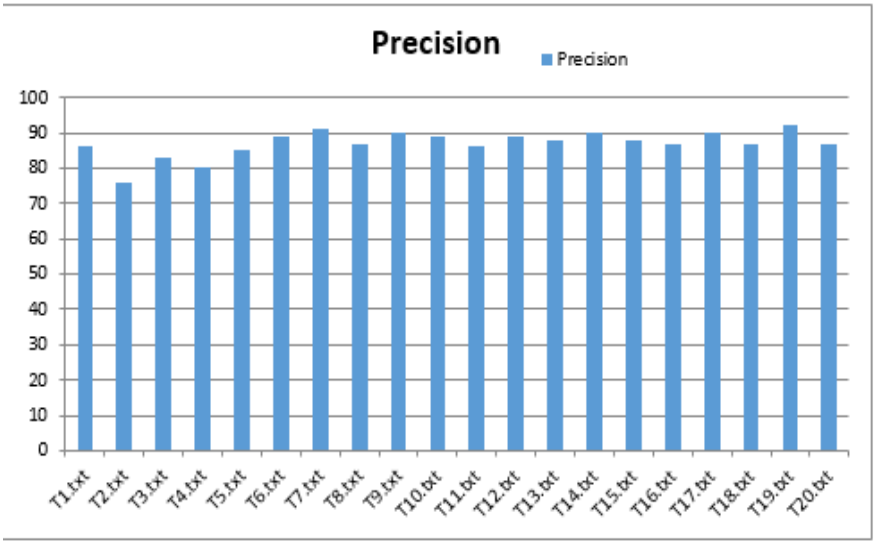



**Fig. 12.** Percentage of Precision Achieved

**Table 2.** Evaluation of System based on input as Web Page

| Input URL (Considering a paragraph in each of the page) | Total no. of words in the document | No. of actual relevant words(TP) | No. of actual irrelevant words(TN) | No.of irrelevant words which turn out to be positive (FP) | No. of relevant words which turn out to be negative(FN) | Recall (%) | Precision (%) | Accuracy (%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 289 | 3 | 285 | 1 | 0 | 100 | 75 | 100 |
| 2 | 43 | 8 | 35 | 3 | 0 | 100 | 73 | 93 |
| 3 | 105 | 10 | 94 | 1 | 1 | 91 | 91 | 98 |
| 4 | 155 | 2 | 153 | 1 | 1 | 67 | 67 | 99 |
| 5 | 200 | 5 | 195 | 2 | 1 | 83 | 71 | 99 |
| 6 | 157 | 14 | 143 | 2 | 0 | 100 | 88 | 99 |
| 7 | 93 | 10 | 83 | 1 | 1 | 91 | 91 | 98 |
| 8 | 54 | 11 | 43 | 1 | 1 | 92 | 92 | 96 |
| 9 | 36 | 15 | 21 | 1 | 0 | 100 | 94 | 97 |
| 10 | 43 | 8 | 35 | 3 | 0 | 100 | 73 | 93 |

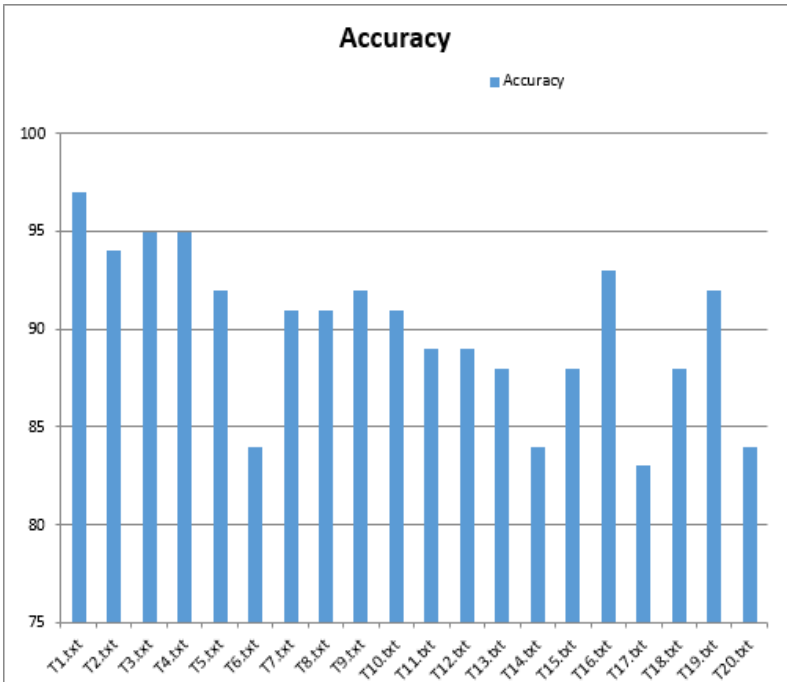



**Fig. 13**. Percentage of Accuracy Achieved

Following is the evaluation of the extracted relevant input entities from the Web page with the help of parameters Recall, Precision and Accuracy along with percentage of these parameters compared right below:

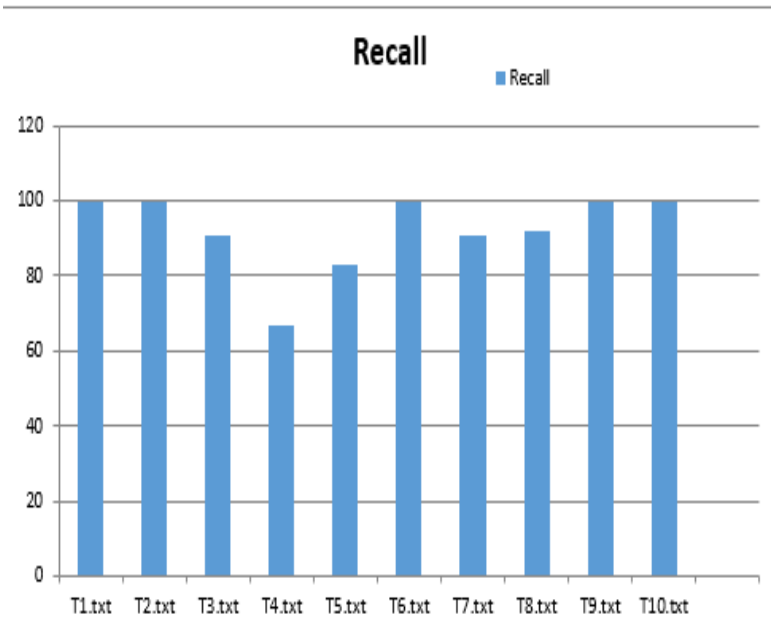



**Fig. 14**. Percentage of Recall Achieved for web page

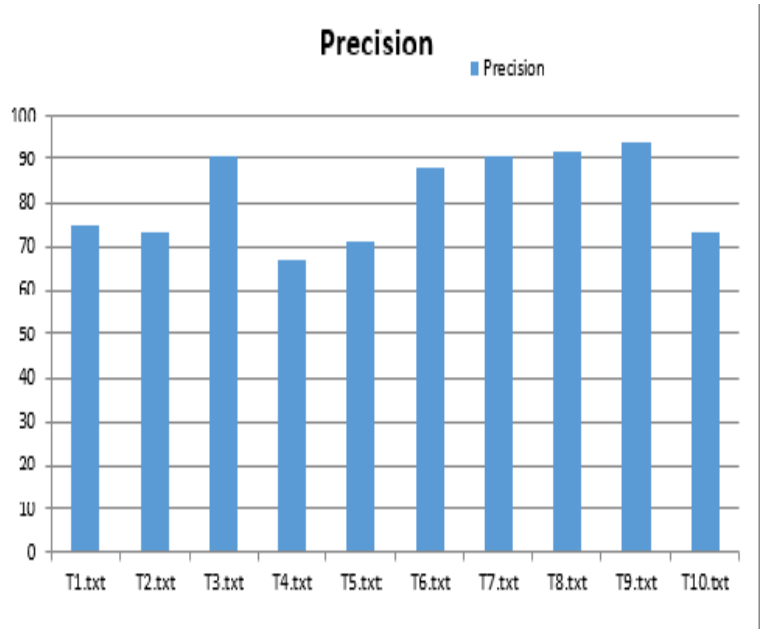



Fig. 15. Percentage of Precision Achieved for web page

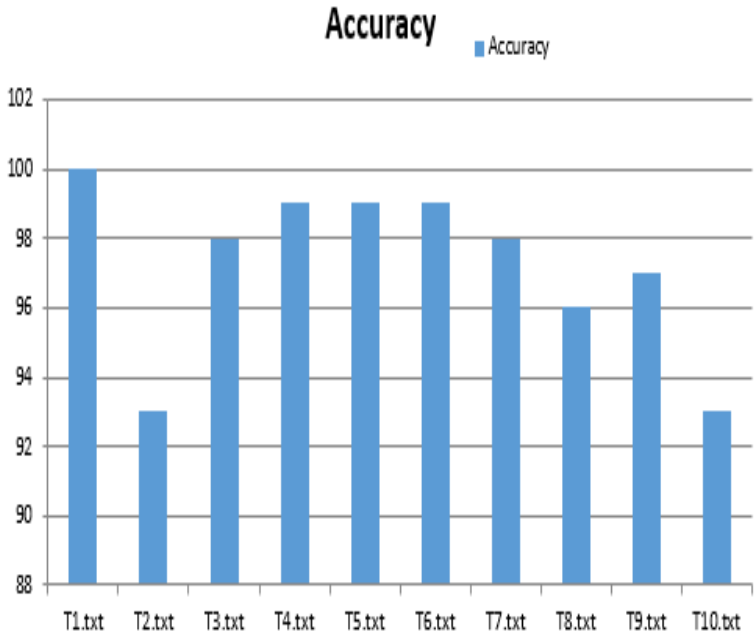



**Fig. 16**. Percentage of Accuracy Achieved for web page

We have done the comparison with previous algorithm which can be seen in the below table, the BHS (Basic HCRF and extended Semi-CRF), NHS (Natural language HCRF and extended Semi-CRF), MHS (Multiple mentions HCRF and extended Semi-CRF), our model gives better results. The evaluation of the various algorithm is based on parameters precision (P), recall (R) and accuracy (F1) compared on the attributes person and location.

The first algorithm is the original HCRF and extended Semi-CRF framework, the BHS (Basic HCRF and extended Semi-CRF) algorithm. The second algorithm is similar to the BHS algorithm. The only difference from BHS is that it adds the natural language features directly into the extended Semi-CRF model. The NHS (Natural language HCRF and extended Semi-CRF) algorithm. The extended Semi-CRF model in the NHS algorithm is trained using both the text nodes from the labelled webpages and the corpus data. The super labels of all sentences from the live search are set as NAME, because we only queried the NAME. The rest of this algorithm is identical to the BHS algorithm. The third algorithm is based on the NHS algorithm. It further adds global multiple mentions feature functions to the HCRF model. We name this algorithm MHS (Multiple mentions HCRF and extended Semi-CRF). These feature functions are all for the business NAME attribute. The summaries of the features at other mentions of the same business NAME candidate are used as feature functions for the current mention. It is some kind of feature sharing.

**Table 3**. Comparison of the system with previous algorithm

| Attributes | Parameters | Person | Location |
|---|---|---|---|
| BHS | P<br>R<br>F1 | 61.8%<br>58.8%<br>60.3% | 100.0%<br>95.6%<br>97.7% |
| NHS | P<br>R<br>F1 | 65.1%<br>47.4%<br>54.9% | 100.0%<br>94.8%<br>97.3% |
| MHS | P<br>R<br>F1 | 65.8%<br>51.1%<br>57.5% | 100.0%<br>94.8%<br>97.3% |
| Information Retrieval System for Web using NLP (The Design System) | P<br>R<br>F1 | 75.1%<br>74.2%<br>74.6% | 100.0%<br>96.4%<br>98.5% |

Thus, we can see from the above comparison table of previous algorithm, the design system gives better results which can be significant when huge of amount of data is taken into consideration.

## 5. CONCLUSION

Web entity extraction plays an important role in Natural Language processing especially in the information retrieval systems. A lot of existing web entity extraction techniques was surveyed. Most of the surveyed literature were DOM tree and database wrappers based methods. Though they produce good results but fail if the web pages are not following the standards of W3C. It also does not work for dynamic web page. The problem of context switching comes with database wrapper based methods. However, little work has been done toward integrating web entity extraction and NLP techniques to achieve better results. The design system is implemented based on the web page layout understanding and text content is analysed through user perception. How a user unconsciously divides the web page by visual perception is incorporated in the system using Visual page segmentation algorithm. This will help in improving user browsing experience. The VIPS techniques along with HCRF for structure understanding and Semi-CRF for text

understanding is used to achieve effective results. The performance of both the models has been boosted by iterative optimization procedure. The decision of Semi-CRF can be used by HCRF for reducing the information redundancy.

Overall accuracy of 86% and 74.6% is achieved for the input text and the web page respectively. The performance as compared to the existing web entity extraction techniques shows a very promising result. The proposed system also works for the web entity extraction for dynamic web page as compared to the existing system where it doesn't work at all. Although the results are achieved for dynamic web page, the time duration with which it is achieved is long. The future work for this system is working on getting improved results based on accurate data and desired time management for the dynamic web pages.

**ACKNOWLEDGEMENTS**

It's my immense honour to write about my guide Dr. Sharvari Govilkar. I am grateful and thankful to have a guide under whom I have been able to carry out this project till the very end. Her constant encouragement and technical expertise have given me right direction to complete this project. She has always been most approachable and also very patient with my queries. Thank you again for all the inputs and leadership you have given me till now.

## REFERENCES

[1] Rini John and Sharvari S Govilkar. Article: Survey of Information Retrieval Techniques for Web using NLP. International Journal of Computer Applications 135(8):23-27, February 2016. Published by Foundation of Computer Science (FCS), NY, USA

[2] Rini John and Sharvari S. Govilkar, " A Novel Approach For Information Retrieval Technique For Web using NLP", International Journal on Natural Language Computing (IJNLC) Vol. 6, No.1, February 2017

[3] 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.

[4] 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.

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

[6] Fedor Bakalov, Bahar Sateli, Ren´e Witte, Marie-Jean Meurs, Birgitta K, "Natural Language Processing for Semantic Assistance in Web Portals", 2012.

[7] R. Witte and T. Gitzinger, "Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients," in 3rd Asian Semantic Web Conference (ASWC 2008), ser. LNCS, vol. 5367. Bangkok, Thailand: Springer, 2008.

[8] Paolo Nesi,Gianni Pantaleo and Marco Tenti, "Ge(o)Lo(cator):Geographic Information Extraction from Unstructured Text Data and Web Documents", in 9th International Workshop on Semantic and Social Media Adaption and Personalization, 2014.

[9] Suma Adindla and Udo Kruschwitz, "Combining the Best of Two Worlds: NLP and IR for Intranet Search", in IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011.

[10] Zhong Liu and Ying Wang, "A Novel method of Chinese Web Information Extraction and Applications", in WASE International Conference on Information Engineering, 2009.

[11] Ruiqiang Guo and Fuji Ren, "Towards the Relationship between Semantic Web and NLP", 2009.

[12] B.Aysha Banu, Dr.M.Chitra , "A Novel Ensemble Vision Based Deep Web Data Extraction" in IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2012.

[13] I.Vijayalakshmi, Sobha Lalitha Devi, "Automatic Information Extraction through Mobile", in ICCCNT'12, Coimbatore, India 2012.

[14] 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.

[15] Z. Nie, F. Wu, J.-R. Wen, and W.-Y. Ma, "Extracting objects from the web", in Proc. 22nd Int. Conf. Data Eng., 2006.

[16] Z. Nie, Y. Ma, S. Shi, J.-R. Wen, and W.-Y. Ma, "Web object retrieval", in Proc. 16th Int. Conf. World Wide Web, 2007.

[17] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma, "Block-based web search", in Proc. Special Interest Group Inf. Retrieval (SIGIR) Conf., 2004.

[18] Zaiqing Nie, Ji-Rong Wen, and Wei-Ying Ma, "Statistical Entity Extraction from the Web" IEEE September 2012.