Information Extraction and Webpage Understanding

M.Sharmila Begum[1], L.Dinesh[2] and P.Aruna[3]

[1] Assistant professor, Department of Software Engineering, Periyar Maniammai University Thanjavur,Tamil Nadu,India

[2] Department of Information Technology, Periyar Maniammai University Thanjavur,Tamil Nadu,India

[3] Assistant professor ,Department of Software Engineering, Periyar Maniammai University Thanjavur,Tamil Nadu,India

Abstract

The two most important tasks in information extraction from the Web are webpage structure understanding and natural language sentences processing. However, little work has been done toward an integrated statistical model for understanding webpage structures and processing natural language sentences within the HTML elements. Our recent work on webpage understanding introduces a joint model of Hierarchical Conditional Random Fields (HCRFs) and extended Semi-Markov Conditional Random Fields (Semi-CRFs) to leverage the page structure understanding results in free text segmentation and labeling. In this top-down integration model, the decision of the HCRF model could guide the decision making of the Semi-CRF model. However, the drawback of the topdown integration strategy is also apparent, i.e., the decision of the Semi-CRF model could not be used by the HCRF model to guide its decision making. This paper proposed a novel framework called WebNLP, which enables bidirectional integration of page structure understanding and text understanding in an iterative manner. We have applied the proposed framework to local business entity extraction and Chinese person and organization name extraction. Experiments show that the WebNLP framework achieved significantly better performance than existing methods.

Keywords: *Natural language processing, Webpage understanding, Information Extraction, Conditional Random Fields*

1. Introduction

The World Wide Web contains huge amounts of data. However, we cannot benefit very much from the large Amount of raw web pages unless the information within them is extracted accurately and organized well. Therefore, information extraction plays an important Role in Web knowledge discovery and management. Among various information extraction tasks, extracting Structured Web information about real-world entities (such as people, organizations, locations, publications, products) Has received much attention of late. However, little work has been done toward an integrated Statistical model for understanding web page structures and processing natural language sentences within the HTML Elements of the web page. Our recent work on Web object Extraction has introduced a template in dependent approach to understand the visually out structure of a webpage and to effectively label the HTML elements with attribute names of an entity. Our latest work on web page understanding introduces a joint model of The Hierarchical Conditional Random Fields (HCRFs) model and the extended Semi-Markov Conditional Random Fields (Semi-CRF's) model to leverage The page structure understanding results in free text Segmentation and labeling. The HCRF model can reflect the structure and the Semi CRF model can make use of the gazetteers. In this top down integration model, the decision Of the HCRF model could guide the decision of the Semi CRF model. However, the drawback of the top-down Strategy is that the decision of the Semi-CRF model could not be used by the HCRF model to refine its decision making. In this paper, we introduce a novel frame work called WebNLP at enables bidirectional integration of page structure understanding and text understanding in an iterative manner. In this manner, the results of page structure understanding and text understanding can be used to guide the decision making of each other, and the

performance of the two understanding procedures is boosted iteratively.

1.1 Overview

The World Wide Web contains huge amounts of data. However, we cannot benefit very much from the large amount of raw web pages unless the information within them is extracted accurately and organized well.Webpage understanding introduces a joint model of Hierarchical Conditional Random Fields (HCRFs) and extended Semi-Markov Conditional Random Fields (Semi-CRFs) to leverage the page structure understanding results in free text segmentation and labeling. In this top-down integration model, the decision of the HCRF model could guide the decision making of the Semi-CRF model. However, the drawback of the top down integration strategy is also apparent, i.e., the decision of the Semi-CRF model could not be used by the HCRF model to guide its decision making.
Our recent work on Web object extraction has introduced a template-independent approach to understand the visual layout structure of a webpage and to effectively label the HTML elements with attribute names of an entity.

In this paper, we introduced the Web NLP framework for webpage understanding. It enables bidirectional integration of page structure understanding and natural language understanding.Specifically, the Web NLP framework is composed of two models, i.e., the extended HCRF model for structure understanding and the extended Semi-CRF model for text understanding. The performance of both models can be boosted in the iterative optimization procedure.The experimental results show that the Web NLP framework performs significantly better than the state-of the- art algorithms on English local entity extraction and Chinese named entity extraction on WebPages.

2. Literature Survey

2.1 Information Extraction

IE technology has not yet reached the market but it could be of great significance to information end-user industries of all kinds, especially finance companies, banks, publishers and governments. For instance, finance companies want to know facts of the following sort and on a large scale: what company take-overs happened in a given time span; they want widely scattered text information reduced to a simple data base. Lloyds of London need to know of daily ship sinkings throughout the world and pay large numbers of people to locate them in newspapers in a wide range of languages. All these are potential uses for IE.

2.2 Empirical Methods in Information Extraction

The first large-scale, head-to-head evaluations of NLP systems on the same text-understanding tasks were the Defense Advanced Research Projects Agency–sponsored Message-Understanding Conference (MUC) performance evaluations of information-extraction systems. Prior to each evaluation, all participating sites receive a corpus of texts from a predefined domain as well as the corresponding answer keys to use for system development. The answer keys are manually encoded templates—much like that capture all information from the corresponding source text that is relevant to the domain, as specified in a set of written guidelines. After a short development phase, the NLP systems are evaluated by comparing the summaries each produces with the summaries generated by human experts for the same test set of previously unseen texts. The comparison is performed using an automated scoring program that rates each system according to measures of recall and precision.

2.3 Extracting Structured Data from Web Page

The World Wide Web is a vast and rapidly growing source of information. Most of this information is in the form of unstructured text, making the information hard to query. There are, however, many web sites that have large collections of pages containing structured data, *i.e.*, data having a structure or a *schema*. These pages are typically generated dynamically from an underlying structured source like a relational database. An example of such a collection is the set of book pages in Amazon. The data in each book page has the same schema, *i.e.*, each page contains the title, list of authors, and price of a book and so on.

2.4 Wrapper Induction Efficiency &Expressiveness

Wrapper is a procedure to extract all kinds of data from a specific web source. First find a vector of strings to delimit the extracted text.

Motivations: hand-coded wrapper is tedious and error-prone. How about web pages get changed? Wrapper induction –- automatically generate wrapper is a typical machine earning technology. Actually we are trying to learn a vector of delimiters, which is used to instantiate some wrapper classes (templates), which describe the document structure free text & Web pages. A good wrapper induction system should be:

Expressiveness: concern how the wrapper handles a particular web site.

Efficiency: how many samples are needed? How much computational is required?

2.5 Wrapper Maintenance Machine Learning Approach

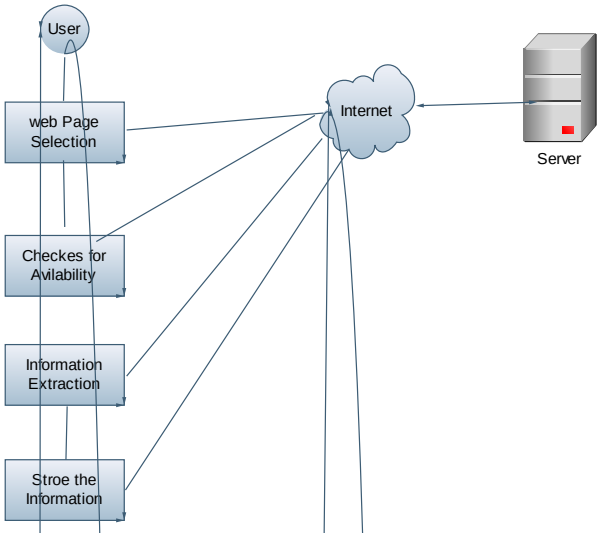
A Web wrapper is a piece of software that enables a Web source to be queried as if it were a database. The types of sources that this applies to are what are called semi structured sources. These are sources have no explicit structure or schema, but have an implicit underlying structure. Even text sources such as email messages have some structure in the heading that can be exploited to extract the date, sender, addressee, title, and body of the messages. Other sources, such as an online catalog, have a very regular structure that can be exploited to extract all the data automatically.

2.6 Hierarchical Wrapper Induction for Semi structured Information Sources

Web pages are intended to be human readable, there are some common conventions for structuring HTML documents. For instance, the information on a page often exhibits some hierarchical structure; furthermore, semi structured information is often presented in the form of lists of tupelos, with explicit separators used to distinguish the different elements. With these observations in mind, we developed the embedded catalog (EC) formalism, which can describe the structure of a wide-range of semi structured documents.

3. Implementation

Extracting information includes these modules to extract the structured data and natural language sentences with in the HTML elements of the webpage.



- ◊ Admin
- ◊ Page Reader
- ◊ Information Extraction
- ◊ Security
- ◊ Information maintenance

3.1 Admin

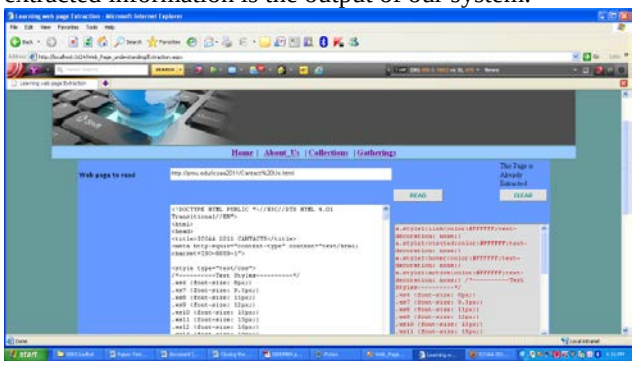
This module provides the facility to control all the operations done by our system. This module completely gives the rights to a single person. This module facilitate the update operation of the data secured by our systemAdmin module is commander module of our system.This module is the central module which integrates other modules. Admin module gives rights to a single person to perform the information extraction.Admin can update the database.Admin only can delete the records from database.

3.2 Page Reader

This module is the heart of our system.This module reads and understands the web pages and its structure.Two algorithms are used by this module.This module provides the facility to read the web page in both directions.Page reader module Reads the enter page source from the server.This module facilitates the source extraction from the server.This produces the output as collection of text information and tags,And also this produces the image links and other links provided in the current site.

3.3 Information Extraction

This is one of the major (Heart) modules of this system.This facilitates the easy ways for information extraction by our system.This module extracts information from the page reader.This module facilitates our system to extract the pure text information from the read source.Then this module extracts the fields and value from the source.This module only extracts exact information.The extracted information is the output of our system.

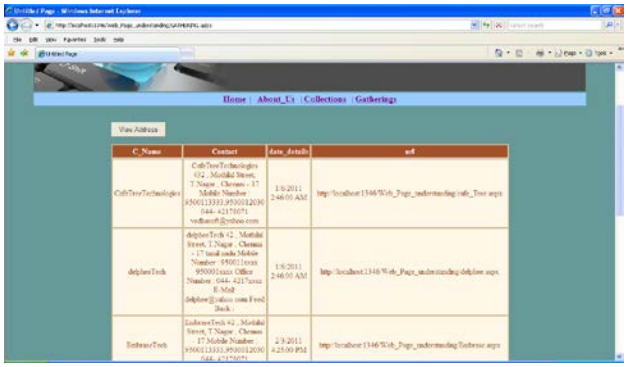


3.4 Security

This module provides the secure operations on our systems.This module allows the secure access by authorized persons only. This module checks for the security verifications like copyrights.This module facilitates our system while understanding the structure of the given input website.This module also restricts from unauthorized access of our application.

3.5 Information maintenance

This is another module of our system.This module provides the facility to store and maintaining the information.It facilitates to store details like Extracted sites and the information extracted from the sites.Information maintenance module facilitates our system for maintaining extracted information by our system.This module provides the facility to store the information. This is also provides the facility to retrieve the information from database



4. Conclusions

Webpage understanding plays an important role in Web search and mining. It contains two main asks, i.e., page structure understanding and natural language understanding. However, little work has been done toward an integrated statistical model for understanding webpage structures and processing natural language sentences within the HTML elements.

In our system, we introduced the WebNLP framework for webpage understanding. It enables bidirectional integration of page structure understanding and natural language understanding. Specifically, the WebNLP framework is composed of two models, i.e., the extended HCRF model for structure understanding and the extended Semi-CRF model for text understanding. The performance of both models can be boosted in the iterative optimization procedure. The auxiliary corpus is introduced to train the statistical language features in the extended Semi-CRF model for text understanding, and the multiple occurrence features are also used in the extended Semi-CRF model by adding the decision of the model in last iteration. Therefore, the extended Semi-CRF model is improved by using both the label of the vision nodes assigned by the HCRF model and the text segmentation and labeling results, given by the extended Semi-CRF model itself in last iteration as additional input parameters in some feature functions; the extended HCRF model benefits from the extended Semi-CRF model via using the segmentation and labeling results of the text strings explicitly in the feature functions. The WebNLP framework closes the loop in webpage understanding for the first time. The experimental results show that the WebNLP framework performs significantly better than the state-of the-art algorithms on English local entity extraction and Chinese named entity extraction on WebPages.

5. References

[1] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma, “Simultaneous Record Detection and Attribute Labeling in Web Data Extraction,” Proc. Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp 494-503, 2006.

[2] Z. Nie, Y. Ma, S. Shi, J.-R. Wen, and W.-Y. Ma, “Web Object Retrieval,” Proc. Conf. World Wide Web (WWW), pp. 81-90, 2007.

[3] J. Zhu, B. Zhang, Z. Nie, J.-R. Wen, and H.-W. Hon, "Webpage Understanding: An Integrated Approach," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 903-912, 2007

[4] Y. Zhai and B. Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.

[5] D. Downey, M. Broadhead, and O. Etzioni, "Locating Complex Named Entities in Web Text," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), pp. 2733-2739, 2007.

[6] O. Etzioni, M.J. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates, "Unsupervised Named- Entity Extraction from the Web: An Experimental Study," Artificial Intelligence, vol. 165, no. 1, pp. 91-134, 2005.

[7] Y. Zhai and B. Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment,"
IEEE Trans. Knowledge and Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.

[8] R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma, "Learning Block Importance Models for Web Pages,"Proc. Conf. World Wide Web (WWW), pp. 203-211, 2004.

[9] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. Int'l Conf. Machine Learning (ICML), pp. 282-289, 2001.

[10] A. Chen, F. Peng, R. Shan, and G. Sun, "Chinese Named Entity Recognition with Conditional Probabilistic Models," Proc. Fifth SIGHAN Workshop Chinese Language Processing, pp. 173-176, 2006.

[11] D. DiPasquo, "Using HTML Formatting to Aid in Natural Language Processing on the World WideWeb,"http://citeseer.ist.psu.edu/dipasquo98using.html, 1998.

[12] C. Jacquemin and C. Bush, "Combining Lexical and Formatting Cues for Named Entity Acquisition from the Web," Proc. 2000 Joint SIGDAT Conf. Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 18

Sharmila Begum received M.E degree in Computer Science and Engineering. She is currently working as a Assistant Professor in Department of Software Engineering in Periyar Maniammai University Thanjavur Tamilnadu India. She has Presented several papers in international conferences and published few papers in PMU journal and published a book named Design and Analysis of Algorithms her research areas are Data Mining, Bio-Medical, OOAD, Networking and Web Programming.

Dinesh received M.Sc degree [5 Years Integrated] in Software Engineering.He is currently pursing his M.E Software Engineering in Periyar Maniammai University Thanjavur Tamilnadu India.

Aruna received MCA and M.Phil degree in Computer Application. She is currently working as a Assistant Professor in Department of Software Engineering in Periyar Maniammai University Thanjavur Tamilnadu India. She has presented several papres in International conferences and her research area is Mobile Adhoc Network.