Wrapper Semi To Structured Database From Multi Web Site Based On Natural Language Processing

[1]Bismar Junatas
[2]Haryanto
[3]Remi Senjaya

[1]Gunadarma University(just_bismar@yahoo.co.id)
[2]Gunadarma University(haryanto@staff.gunadarma.ac.id)
[3]Gunadarma University(remi@staff.gunadarma.ac.id

Abstract

The number of data source on internet has increased in volume and type since the last decade, causing problems to query the data or information because of the diversity, dynamic and heterogeneity of the data source or information. Therefore, to simplify the task of obtaining information, several tools have been created for extracting the data from multiple web sources, including Wrapper. Wrapper facilitates the access to Web-Based information sources by providing a uniform querying and data extraction capability. It consists of a set of extraction rules and the code required to apply the rules in order to make the wrapper extracts the right and specified information. The research focuses on how to query the data of rooms and rates hotels in Indonesia by proposed a single wrapper which will change the semi data to structured database based on Natural Language Processing.

Keywords : Information Extraction, Multi Web Site, Natural Language Processing, Wrapper

1 Introduction

For many activities that run nowadays, it desperately needs of a variety of information such as for decision-making, planning, evaluation and so forth. The sources of information has become increasingly diverse and more than a million at the time because of the improvement on information technology and the internet, makes no geographic boundaries and time for exchanging and generating the diversity of information sources.

The information source that available today (which is contained in a web) has a different method of presenting the information even though the purpose is same or similar with another. Obviously, this caused by a high level of autonomy on web technology and it makes the need of different methods to interpret each web site although in the end it will have the same result. Moreover, accessing many heterogeneous sources also needs appropriate concepts in query to get the right respond. So, the considered points of the problems in this work are: (i) how is the way of the automation process about dump data, (ii) how to access multi web site for extracting semi-structured data to become structured database, and (iii) how to make multisource of data representation become into single view representation. The research focuses on extraction data from multi web site, especially Indonesian local web sites, by using a wrapper. In the end of the research, it proposes a single wrapper program that can extract semi-structured data and change it into structured database from multi web site, which based on Natural Language Processing (NLP). Hopefully the research can be useful for everyone who reads it.

2 Theory

2.1 Information Extraction

The presence of information will affect people to acquire knowledge while doing their activities. People also can obtain and extract the information easily, either through newspapers, television, radio, and even web sites. The diversity, heterogeneity, and the size of information generated through the internet, making the challenge how to get the right information in accordance with user needs. The challenge has created methods, terms, technologies, sciences, or systems, to standardize and integrate the data or information which will be collected in accordance with user needs. One of the technologies is Information Extraction.

Information Extraction is the name given to any process which selectively structures and combines

data which is found, explicitly stated or implied, in one more texts [6]. The final output of the extraction process varies; in every case, however, it can be transformed so as to populate some type of database. Another definition for Information Extraction (IE) acknowledged in [18]. It acknowledged that IE is the process of identifying the particular fragments of an information resource that constitute its core semantic content. A number of IE systems have been proposed for dealing with free text and semi structured text.

Generally, these definitions have represented the IE in overall and become the basic definition for all understanding of Information Extraction.

2.2 Information Extraction in Context

Information Extraction (IE) differs from Information Retrieval (IR) and Natural Language Understanding (NLU) [3]. Typically, IR involves searching and retrieving from a collection of documents a subset which is relevant to a query in terms of keyword matching. The functionality of NLU is hard to characterize and evaluate, but usually it is more sophisticated. There is no clearly-cut boundary among IE, IR, and NLU. However, IE can be viewed as a task that lies between IR and NLU from the perspective of the complexity of functionality required.

3 Approach

3.1 Wrapper

To extract information from semi-structured information resources, information extraction systems usually rely on extraction rules tailored to the source, generally called Wrapper. Wrapper is software module that helps to capture the semi-structured data on the web into a structured format [18].

In the database community, a wrapper is a software component that converts data and queries from one model to another. In the Web environment, its purpose should be to convert information implicitly stored as an HTML document into information explicitly stored as a data-structure for further processing. Wrapper facilitates access to Web-based information sources by providing a uniform querying and data extraction capability.

Refers to [18], a wrapper has three main functions, such as:

- Download: It must be able to download HTML pages from a web site.
- Search: Within a resource it must be able to search for, recognize, and extract specified data.
- Save: It should save the data in a suitably structured format to enable further manipulation. The data can then be imported into other applications for additional processing.

So, to support the functions, the wrapper consists of a series of rules and some code to apply the rules and generally speaking about specific case to the source [8]. It is specialized to a single information source, and since a semi-structured web source normally presents its contents to the browser in a uniform way, a single wrapper is enough for each web site.

3.2 Classification of Wrapper

Furthermore, the techniques of wrapper generation can be classified into three categories, as follows:

- Single Wrapper In Single Wrapper classification, the extraction rules are written by programmer through careful examinations of a set of sample pages. The programmer must describe a new extraction rule for each different web site since the structures of each web are different. For examples: Hotel A, B, and C are the best hotel in Jakarta. Each hotel has a web site. And for extracting data, it might each web has own wrapper for gained that (Wrapper hotel A, wrapper hotel B, and wrapper hotel C). For more detail, take a look on figure 1.

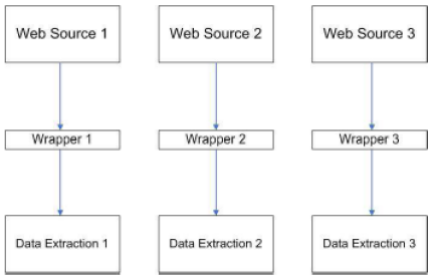

Figure 1: Single Wrapper Architecture

Figure 1 represents the architecture of single wrapper. And as acknowledged previously, for each web sources has its own wrapper for data extraction. In other means, the development of the wrapper depends on the structure of each web sources.

- Generic Wrapper

  Generic Wrapper classification is enhanced of the previous wrapper classification (single wrapper). It allows user does not have to write extraction rules for each web site. It only has one wrapper program that can extract from multi web site. It also reduces the limit of user to obtain comparable information from the known information sources. Figure

2 represents the architecture of Generic Wrapper, the enhanced of single wrapper model. As acknowledged previously, in generic wrapper, its only has one wrapper for several web sources for data extraction, reducing the time cost for data extraction from the previous classification.

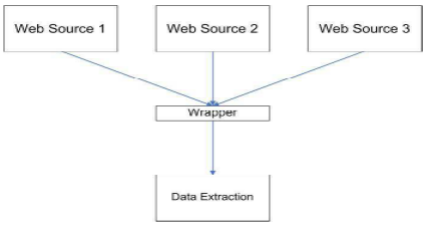

Figure 2: The Generic Wrapper Architecture

- One of the existing wrapper programs that used this classification is W4F (WysiWyg Web Wrapper Factory). W4F is toolkit for the generation of wrappers of web sources. W4F consists of a retrieval language to identify Web sources, a declarative extraction language (the HTML Extraction Language) to express robust extraction rules and a mapping interface to export the extracted information into some user-defined data-structures

- Smart Wrapper

  Smart Wrapper classification is the fully automated wrapper program that cooperate single and generic wrapper in one performed and also the development of both wrapper classification. It just has one rule that can extract and update data extraction from multi web site.

  This wrapper is a perfect wrapper. But to create wrapper which based on this classification, it needs free access for all web sites that exists in the World Wide Web. Means there are no security questions if user wants to get specified information from the web which he opens. It also needs the structure of all the web sites has the same structure. That is why until the research has been done, the implementation of this wrapper classification has not been found.

3.3 Natural Language Processing

Natural Language Processing (NLP) is subject of natural language are formalism, algorithms and methods of computer linguistics [21]. Automatic translations, Speech recognition, Spelling and Grammar checking, Dialogue Systems are examples of Natural Language Processing. NLP is experiencing rapid growth as its theories and methods are deployed in a variety of new language technologies. Natural language also understanding systems convert samples of human language into more formal representations such as parse trees or first order logic that are easier for computer programs to manipulate.

And for this reason, it is important for a wide range of people to have a working knowledge of NLP. Within industry, it includes people in human-computer interaction, business information analysis, and Web software development. Within academia, this includes people in areas from humanities computing and corpus linguistics through to computer science and artificial intelligence.

Furthermore, the study of language is part of many disciplines outside of linguistics, including translation, literary criticism, philosophy, anthropology and psychology. Many less obvious disciplines investigate language use, such as law, hermeneutics, forensics, telephony, pedagogy, archaeology, cryptanalysis and speech pathology. Each applies distinct methodologies to gather observations, develop theories and test hypotheses. Yet all serve to deepen our understanding of language and of the intellect that is manifested in language.

3.4 General Development Methods of Wrapper using NLP

There are three main processes are done in natural language processing, such as syntactic analysis, semantic interpretation and contextual interpretation. Syntactic analysis or parsing is the process of determining the structure of a sentence based on a particular grammar and lexicon. Parsing can be done in a top-down and bottom-up, each has advantages and disadvantages. Top-down parsing cannot handle the grammar with left-recursion, while the bottom-up parsing cannot handle the grammar with empty production. Because it is the best method of parsing that can combine both these ways.

Semantic interpretation is the process of translating a sentence into a common form of representation called a logical means to form regardless of context. Two main processes are needed in forming the logical form is to determine the role of each word and phrase in a sentence, and the selection of the proper meaning of words to form sentences that make sense. The role of words and phrases in a sentence can be represented in the form of predicate-argument to use common or thematic roles. While the process of selecting the correct meaning of words can be done with the selection of restrictions or context activation.

The method used in the process of wrapping is a method of parsing performed word for word taken from the web page downloads. This web page is downloaded and which are parsed only

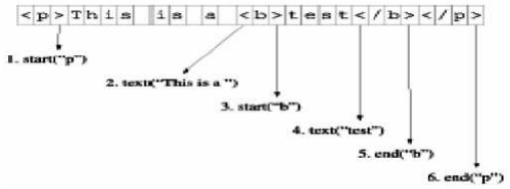

Figure 3: Event-Oriented Parsing from HTML

web page text and text / html which is the MIME type. To parse HTML pages used event-oriented parser. Where event-oriented parser does not build the structure of the document, however it is to generalize the function, as shown in figure 3.

During the parsing process of duplication is based on the content page, but links from pages found in duplicate will be ignored because it will reduce the bandwidth. This parser will not generate any HTML tags. But selecting only the document structure, logical format, and physical formats such as bold and italics. Information about the colors, backgrounds, type fonts, etc. will be ignored. Authors and Affiliations

3.5 Research Position

For this research, according to the research that has been done for developing a wrapper, Single Wrapper is the right or suitable classification in developing the new wrapper program for extracting the data from Indonesian local web sites.

4 Experiment

4.1 The Scenario of running The Wrapper

Firstly, the scenario is to measure the parameters of rooms and rates from the origin sites (Bali Dynasty Hotel web site, Jogjakarta Plaza Hotel web site, Oasis Amir Hotel web site) then continues to measure the parameters of rooms and rates from the wrapper program that has been created. The way to measure the parameters is to look at the rooms and rates from each hotel sites and the wrapper program. In this case, as a reference measurement is the original sites, and as an object to be tested is a wrapper program that has been made.

Then after these scenarios running and finishes, the next activity is about to compare the results between the wrapper program and each hotel web sites. The goal is to prove whether this system is said to be worth supporting to be one of wrapper program and to see whether the results from the program and web sites are same or different.

4.2 Results

There are three wrappers for three web sites in this thesis. After each wrapper has been being run simultaneously every 5 day, they have given results to author. This section explains the result of testing program that will be explained in the following way:

Table 1: Result from Bali Dynasty Hotel

Time	Error(%)
August 21, 08.00	0
August 22, 09.00	0
August 23, 10.00	0
August 24, 11.00	0
August 25, 12.00	0
August 26, 13.00	0
August 27, 14.00	0
August 28, 15.00	0
August 29, 16.00	0
August 30, 17.00	0

Table 2: Result From Jogjakarta Plaza Hotel

Time	Error (%)
August 21, 08.00	0
August 22, 09.00	0
August 23, 10.00	0
August 24, 11.00	0
August 25, 12.00	0
August 26, 13.00	0
August 27, 14.00	0
August 28, 15.00	0
August 29, 16.00	0
August 30, 17.00	0

Table 3: Result from Oasis Amir Hotel

Time	Error (%)
August 21, 08.00	0
August 22, 09.00	0
August 23, 10.00	0
August 24, 11.00	0
August 25, 12.00	0
August 26, 13.00	0
August 27, 14.00	0
August 28, 15.00	0
August 29, 16.00	0
August 30, 17.00	0

Refers to the table 1, table 2, and table 3, no error appears after the wrappers have been run and finish the task. Then the results section will continue with the result of merging all data extraction into single view representation.

Table 4: Result of Merging the Data

Time	Error (%)
August 21, 08.00	0
August 21, 08.00	0
August 21, 08.00	0
August 22, 09.00	0
August 22, 09.00	0
August 22, 09.00	0
August 23, 10.00	0
August 23, 10.00	0
August 23, 10.00	0
August 24, 11.00	0
August 24, 11.00	0
August 24, 11.00	0
August 25, 12.00	0
August 25, 12.00	0
August 25, 12.00	0
August 26, 13.00	0
August 26, 13.00	0
August 26, 13.00	0
August 27, 14.00	0
August 27, 14.00	0
August 27, 14.00	0
August 28, 15.00	0
August 28, 15.00	0
August 28, 15.00	0
August 29, 16.00	0
August 29, 16.00	0
August 29, 16.00	0
August 30, 17.00	0
August 30, 17.00	0
August 30, 17.00	0

Refers to the table 4, it shows that the process of merging data from several sites works well because no error appears during the wrapper process. The analysis of the wrapper results will be explained afterwards in new section.

4.3 Table Explanation

This section explains contains of the table, such as Time, Hotel, and Error, as follows:

1) *Time*: TimeTime is the variable when the wrapper has being run simultaneously. The time includes day, date, and time.

2) *Hotel*: Hotel is the variable that mentioned the hotel's name, which means the hotel's name contains Bali Dynasty Hotel, Jogjakarta Plaza Hotel, and Oasis Amir Hotel.

3) *Error*: Error is the variable that mentioned how many times the data result of wrapper does not appropriate with the data original from the sites.

An equation below will explain how to count the error

$$error(percent) = (\frac{totaldataerror}{totaldata})x100percent$$

Refers to the equation, there are three variables, such as: total data, total data error, and error in percent. Total data means the amount of data that are being extracted by the wrapper, total data error means how many data that has been extracted by wrapper does not appropriate with the original, and the last is error in percent which means the calculation of data error in percent.

For example, total data rooms and rates from Bali Dynasty Hotel Web are eight and those data has been extracted by the wrapper and has been stored into file text. After the comparison between data extraction result and data original from web, two data has not appropriated. So the calculation of error is two divided by eight and after then multiply by one hundred percent, which equals twenty five percent.

4.4 Analysis

All of the processes, according to the table results, are working well. No error appears after the process has finished. So, it can be concluded that the single wrapper program that has been developed is a good development. Starts from the tagging process of each site until the merging process, no error appears in the result.

5 Conclusion and Future Work

5.1 Conclusion

To simplify the task of obtaining information from the vast number of information sources, several tools has been created for extracting the data from multiple web sources, including Wrapper. Wrapper consists of a set of extraction rules and the code which required applying the rules. So it depends on the programmer to write the necessary grammar rules in order to make the Wrapper extracts the right and specified information. The wrapper which developed in the research surely can help users to get specified information from multi web site based on Natural Language Processing. So, even the structures of Indonesian local web sites are different between one to another site, but obviously we can make a kind of wrapper which can extract specified information from the sites. And for overall, Users can make multi source of data representation becomes single view representation.

5.2 Future Work

Furthermore, the wrapper really needs to upgrade. It is still to write different codes for several target sources in order to get the information or data which contains on the target sources. So, in the future work, the wrapper program can be enhanced

to become generic wrapper, or maybe smart wrapper. And it is very obviously that users will use the wrapper as a tool to get the appropriate information since they always need the information which available online in the internet or web site

References

[1] N. Ashish, and Knoblock, "Wrapper generation for semi-structured internet sources," Information Sciences Institute and Department of Computer Science, University of Southern California, 2000.(references)

[2] R. Baumgartner, W. Gatterbauer, and G. Gotlob, "Web Data Extraction System," 2007, unpublished.

[3] F. Chen, "Learning information extraction patterns," Iowa State University, 2000, unpublished.

[4] K. L. Clark, and T. W. Hong, "Towards a universal web wrapper," unpublished.

[5] J. Cogliati, Non-Programmers Tutorial for Python, 2002.

[6] J. Cowie, and Y. Wilks, "Information Extraction", 1996, in press.

[7] S. Farrar, and S. Moran, "The e-linguistics toolkit," 2006, unpublished.

[8] G. Fiumara, "Automated information extraction from web sources: a survey," Dipartimento di Fisica, Universit'a degli Studi di Messina, 2008, in press.

[9] G. Gotlob, R. Baumgartner, and S. Flesca, "Supervised wrapper generation with Lixto," VLDN Conference, 2001, in press.

[10] A. Grosskurth, and M. W. Godfrey, "A reference architecture for web browsers," Canada, 2003, unpublished.

[11] P. W. Handayani, "Automated tag generation based on context analysis," Germany, 2008, unpublished.

[12] C. –N. Hsuy, C. –H. Changz, and C. –H. Hsieh, "Reconfigurable web wrapper agents for biological information integration," Taiwan, 2002, unpublished.

[13] J. Liu, N. Zhong, and Y. Yao. Web Intelligence. Springer, 2003.

[14] X. Liu, "Natural language processing," unpublished.

[15] E. Loper, E. Klein, and S. Bird, "Natural Language Processing," 2008, unpublished.

[16] A. Manggolo. Modul Pengenalan Pemrograman Python. 2003.

[17] G. A. Mihaila, "Websql-an sql-like query language for World Wide Web," 2008, unpublished.

[18] R. Mohapatra, "Information extraction from m dynamic web sources," 2004, unpublished.

[19] P. Nakov, A. Schwartz, and B. Wolf, "Supporting annotation layers for natural language processing," 2005, unpublished.

[20] S. Neli, "Internet data acquisition, , search and processing," Auburn University, 2009, unpublished.

[21] A. Pottharst, "Natural language processing introduction and motivation," 2008, unpublished.

[22] B. Rahardjo, "Aplikasi web grabber untuk mengambil halaman web sesuai dengan keyword yang diinputkan," Petra Christian University, 2004, in press.

[23] M. T. Roth, and P. Schwarz "A wrapper architecture for legacy data sources," 2001, in press.

[24] I. W. S. Wicaksana, "A peer to peer (p2p) based semantic agreement approach for spatial information interoperability," Gunadarma University, 2006, unpublished

[25] J. Carme, M. Goebel, and M. Cheresna. Web Wrapper Using Compound Filter Learning. Book.