Web Mining: A Brief Survey on Data Extraction Techniques

Vijay Bagdi
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur

Sulabha Patil
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Nagpur

Abstract
In last few years, we faced problem regarding extraction of data from web pages. In this paper we proposed to address problem of web data extraction techniques related to areas such as natural language processing, language and grammar, machine learning, information retrieval and Ontologies. As consequence they represent very distinct feature and capabilities which make direct comparison difficult to be done.

Keywords
Ontology, Web Data Extraction, Web Crawler, Road Runner

1. INTRODUCTION
With expansion of internet, the wealth of data regarding many subjects has become online. Usually, browsing and keyword are intuitive form of searching data on web however such strategies have several limitations. Keyword searching is more efficient than browsing but often returns huge amount of data beyond user's requirement while browsing is not suitable for locating particular items of data.

Some researchers have push ideas to manipulate data from database area. Yet, most web data is unstructured or semi structured and cannot be manipulated using traditional techniques. To overcome this problem, possible strategy is need to implement to extract data from web sources to populate database.

The traditional approach for extracting data from web sources is to write specialized program called wrappers which identify data of interest and map them to some suitable format. Recently many techniques to proposed to better address the issue of generating wrappers form data extractions [3, 4, 9, 10, 11, 12 15, 17, 19, 22, 25, 29, 31, 32]. Such techniques are based on several techniques such as declarative languages [4, 10, 17], HTML structure analysis [11, 25, 32], Natural Language Processing [15, 29, 33], machine learning [9, 19, 22], data modeling [3, 31] and Ontologies [12].

The problem of generating wrapper from web extraction can be stated as follows. Given web page S containing a set of implicit objects, determining mapping W that populates data repository R with objects in S. the mapping W must be capable of recognizing and extracting data from any other page S' similar to S. we use term similar in very empirical sense, meaning pages provided by same site. In this context, wrapper is program that executes the mapping W. A common goal of wrapper generation tool is to generate wrappers that are highly accurate and robust, while demanding little effort from wrapper developer.

As more and more techniques appear for data extraction from web, the need arise for analysis of their capabilities and features.

2. OVERVIEW OF WEB DATA EXTRACTION TECHNIQUES

2.1 Markov Logic Networks
Markov Logic Networks (MLN's) [36] provide powerful probabilistic modeling framework based on first order logic. Formally, MLN is set of pairs (F_i, w_i), where F_i is first order formula and w_i is corresponding weight. The weight of formula is essentially a measure of its importance. Formulas are specified over set of application specific predicates. The set of predicates are categorized as query (or hidden) and evidence (or observed) predicates, and formulas capture various relationship between these predicates.

For example, in our extraction application, the query predicates are the attribute labels assigned to page nodes n like IsName(n), IsAddress(n), etc., and the evidence predicates are the observed content and structural features over nodes like Has5Digits(n), FirstLetterCapital(n), Close(n1,n2), etc.Then using such predicates, formulas like 8n Has5digits(n) IsZipCode(n) and 8n1; n2 IsName(n1) ^ IsAddress(n2) Close(n1; n2) are formed.

Now, for a web site W, let x be the set of evidence predicates that are true for pages in W.

Thus the inference problem is an instance of the weighted MAX-SAT problem, which is known to be NP-hard [12]. An ancient approach to this problem is stochastic local search, exampled by the MaxWalkSAT solver [38].

2.2 Structured Data Extraction Techniques:

2.2.1 WEB CRAWLER
It is nothing but a computer programs that browses through the World Wide Web in amethodical fashion. The process is called crawling or spidering. Mostly used in search engines. There are two types in Web Crawler those are External and Internal Web Crawler. External crawler orders unknown websites and initiates an internal crawl on the first page of the website only. Internal crawler will only crawl through internal pages of the websites returned by the external crawler. The web pages generated by the internal crawlers are more reliable. Web

Crawlers are a central part of search engines, and details on their algorithms and architecture are kept as business secrets. When crawler designs are published, there is often an important lack of detail that prevents others from reproducing the work. There are also emerging concerns about "search engine spamming", which prevent major search engines from publishing their ranking algorithms. Web crawlers typically identify themselves to a Web server by using the User-agent field of an HTTP request. Web site administrators typically examine their Web servers' log and use the user agent field to determine which crawlers have visited the web server and how often. The user agent field may include a URL where the web site administrator may find out more information about the crawler. Spam bots and other malicious Web crawlers are

unlikely to place identifying information in the user agent field, or they may mask their identity as a browser or other well-known crawler. It is important for Web crawlers to identify themselves so that Web site administrators can contact the owner if needed. In, crawlers may be accidentally trapped in a crawler trap or they may be overloading a Web server with requests, and the owner needs to stop the crawler.

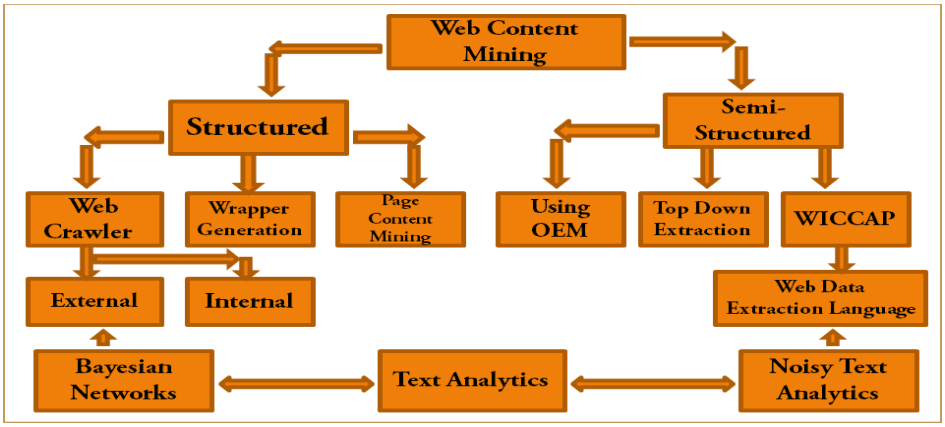

Figure 1: Web Content Mining

Identification is also useful for administrators that are interested in knowing when they may expect their Web pages to be indexed by a particular search engine.

Querying the data in the source requires generating an appropriate URL for the document that contains the answer. Constructing the URL is quite different compared to translation from a set of standard operators to another set of operators native to a database-like source? The answers are not strictly typed objects in some standard representation. Instead, the answers have to be extracted from the semi-structured data stored in HTML documents, and these documents may contain data that is irrelevant to the query. The structure of these documents may be volatile and this affects the extraction process. Domain knowledge about the data source is also embedded in HTML documents.

2.2.2 WRAPPER GENERATION

When you query for some information on the web, this technique works on the pages that are already ranked by traditional search engines. A Web accessible source typically does not support a schema, nor does it support a set of operators such as the set of standard (relational) operators or operators supported by an IR search engine. Querying the data in the source requires generating an appropriate URL for the document that contains the answer. Constructing the URL is quite different compared to translation from a set of standard operators to another set of operators native to a database-like source? The answers are not strictly typed objects in some standard representation. Instead, the answers have to be extracted from the semi-structured data stored in HTML documents, and these documents may contain data that is irrelevant to the query. The structure of these documents may be volatile and this affects the extraction process. Domain knowledge about the data source is also embedded in HTML documents and must be extracted and in addition to answering queries, the wrapper will provide information on the capability of the sources.

2.3 Semi - Structured Data Extraction Techniques:

2.3.1 OEM (OBJECT EXCHANGE MODEL) AND SCHEMA

2.3.2

Knowledge:

This mining is one of the best technique in Semi-Structured Data Extraction Techniques. In this approach, relevant information is extracted by embed in a group of useful information and storing it Object Exchange Model (OEM) and Schema knowledge mining can then be applied on this information. It helps the user to understand the information structure on the web more accurately. OEM, that believe is well suited for information exchange in heterogeneous, dynamic environments and it is flexible enough to encompass all types of information, yet it is simple enough to facilitate integration and also it includes semantic information about objects. A main feature of Object Exchange Model is self describing, we need not define in advance the structure of an object, and there is no notion of a fixed schema or object class.

2.3.3 TOP DOWN EXTRACTION:

It extracts complex objects from a set or rich web sources and converts it into less complex objects. This helps in extracting object attributes and building a good web structure. To extract the data, we need some description of what to extract. A common approach to providing such description is to build a specific grammar which details the surroundings of each piece of data to extract In this work, however, we consider a new approach in which the description of what to extract is fully based on a small set of examples provided by the user. Based on this idea, we propose a top-down strategy that extracts complex objects decomposing them in objects less complex, until atomic objects have been extracted. Through experiment, we demonstrate that just a couple of examples are sufficient for extracting hundreds of objects from new Web pages. Our approach is simple, intuitively appealing, quite effective, and does not suffer from them in drawbacks of alternative approaches in the literature. We discuss here our top-down strategy to extract complex objects from data rich Web sources.

Before discussing the details of the proposed strategy, we define some basic concepts and terminology required to deal with the hierarchical structure of complex objects. Since the

structure of Web page objects is not always flat, it is necessary to introduce the concept of a complex object with a hierarchical structure. The idea is to collect a couple of example objects from the user and to use this information to extract new objects from new pages or texts. We propose a top-down strategy that extracts complex objects decomposing them in objects less complex, until atomic objects have been extracted. Through experimentation, With regard to user involvement, three principal approaches for wrapper generation can be distinguished:

(1) Manual wrapper programming, in which the system merely supports a user in writing a specific wrapper but cannot make any generalizations from the examples provided by the user;

(2) wrapper induction, where the user provides examples and counterexamples of instances of extraction patterns, and the system induces a suitable wrapper using machine learning techniques.

(3) semi-automatic interactive wrapper generation, where the wrapper designer not only provides example data for the system, but rather accompanies the wrapper generation in a systematic computer-supported interactive process involving generalization, correction, testing, and visual programming techniques.

2.3.4 WEB DATA EXTRACTION LANGUAGE

The other Semi-Structured Data Extraction Technique and it is primary responsibility is to convert web data to structured data sensitive to user requirement and It stores data in the form of tables since most data is semi-structured document that are easy to be rendered by the browser and read by the user. Web data extraction systems in use today transform semi-structured Web documents and deliver structured documents to end users. Some systems provide a visual interface to users to generate the extraction rules.

However, to end users, the visual effect of Web documents is lost during the transformation process.

2.4 HTML – aware tool

W4F (World Wide Web Wrapper Factory). W4F [32] is toolkit for building wrappers. W4F divides wrapper development process in three phases : first, the user describes how to access the document, second he describes what pieces of data to extract and third he declares what target structure to use for storing the data extracted. A document is first retrieved from the web according to one or more retrieval rules. Once retrieved it is fed to HTML parser that constructs a parsing tree following the Document Object Model (DOM) [35]. Users can write extraction rules for locating data into parsing tree. The extracted can be stored using W4F internal format called NSL (Nested String List). NSL structures are exported to upper level applications according to specific mapping rules. The language define by W4F to define extraction rules is called HEL (HTML Extraction Language). An extraction rule is assignment between variable name and path expression.

RoadRunner– A recent tool that further explores the inherent features of HTML document to automatically generated wrappers is RoadRunner [11]. It works by comparing HTML structure of two (or more) given sample pages belonging to same page class, generating as a result a schema for data contained in the pages. From this schema a grammar is inferred which is capable of recognizing instances of the attributes identified for this schema in sample pages (or pages in the same class). To accurately capture all possible structure variations occurring on pages of same page class, it possible to provide more than two sample pages. All extraction process is based on algorithm that compares the tag structure of the sample pages and generates regular expressions that handles regular structural mismatches found in between two structures.

2.5 Ontology

This approach is mainly represented by the works of data extraction group [12] at Brigham Young University. In this technique, ontologies are previously constructed to describe the data of interest, including relationship, lexical appereance and context keywords. By parsing this ontology, the tool can automatically produce database by recognizing and extracting data present in documents or pages given as input. Prior to the application of ontology, the tools require the application of an automatic procedure to extract chunks of text containing data items (or records) of interest.

3. CONCLUSION

In this paper we presented short survey of existing techniques for generation of wrappers to extract data from web sources. In all above techniques, the main goal is to ease the task of wrapper development, which is traditionally accomplished by code writing.

We also analyze the qualitatively techniques by examining how they support some features that we consider important to accomplish the generation of wrappers and data extraction process performed by them.

4. REFFERENCES

[1] Abascal R. And Sanchez, J.A. Xtract : structure Extraction from Botanical Textual Description. In proceeding of String Processing and Information Retrieval Symposium and International Workshop on Groupware SPIRE/CRIWG (Camcum, Mexico, 1999).

[2] Abiteboul .S. Querying Semi-structured data. In Database Theory- ICDTS'97 – 6thInternational Conference , Delphi, Greece, January 8-10, 1997.F.N. Afrati and P. Kolaitis. Eds. Vol. 1186 , Lecture Notes in Computer Science.

[3] Adelberg B. Nodose . A tool for Semi-Automatically Extracting structured and Semi-Structured Data from Text Document, SIGMOD Record 27, 2(1998).

[4] Arocena . G.O. And Mendelzon A. O. WebOQL : Restructuring Documents, Databases and Webs. In Proceeding of 14th IEEE International conference on Data Engineering (Orlando, Florida, 1998)

[5] Baumgartner R, Flesca S. And Gottlob G. Visual Web information Extraction with Lixto. In proceeding of 26th International Conference on Very Large Database Systems (Rome, Itly, 2001).

[6] Bray.T., Pauli.J. And Mcqueen M.S Extensible Markup Language (Xml) 1.0

[7] Brin S., Motwani R., Page L. And Winograd T. What Can You Do With Web In Your Pocket? Data Engineering Bulletin 21,2 (1998)

[8] Buneman P. Semi-structured data. In proceedings of the sixteenth ACM-SIGACT-SIGMOD-SIGART . Symposium on Principles of Database Systems (Tueson, Arizona, 1997)

[9] Califg M. E. And Mooney R. J. Relational Learning of Pattern-Match Rules for Information Extraction.In proceeding of Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on

Innovative application of Artificial Intelligence (Orlando, Florida, 1999).

[10] Crescenzi V. And Mecca G. Grammars have Exceptions, Information Systems 23,8.

[11] Crescenzi V, Mecca G., Nad Merialdo P.Roadrunner : Towards automatic data extraction from web sites. In proceeding of 26th International Conference on Very large Database Systems (Rome, Itly, 2001)

[12] Embley D. W., Campbeli D. M. , Jiang Y. S. , Liddle S. W. , Kaing Y. , Quass D. And Smith R. -D. Conceptual-Model-Based Data Extraction from Multi-Record web pages. Data and Knowledge Engineering 31, 3 (1999).

[13] Embley D.W., Jiang Y. S. And Ng. Y. K. Record Boundary Discovery in web documents.In proceedings ACM SIGMOD International Conference on Management of Data (Pholadephia, Penssylvania, USA, 1999).

[14] Florescu D., Levy. A. Y. And Medelzon, A. O. Database Techniques for the World-Wide-Web: a Survey SIGMOD record 27, 3 (1998).

[15] Freitag D.Machine Learning for Information Extraction in Informal Domains. Machine Learning 39, 2/3 (2000)

[16] Golgher P. B., Di Silva A. S. , Laender A. H. F. And Ribiero Neta. B. A. Bootstrapping for example based Data extraction. In proceedings of tenth ACM International Conference on Information and Knowledge management (Atlanta, Georgia, 2001)

[17] Hammer J. , Garcia-Molina, H. Nestorov, S. Yerneni, R., Breunig M. And Vassalos V. Template-base wrappers in TSIMMIS system. SIGMOD Record 26, 2(1997)

[18] Hammer J., Mchugh J. And Garcai-Molina H., Semi-Structured Data: The Tismmis Experience. In proceedings of the first East-Europian Symposium on Advances in Database and Information Systems.(ADBIS'97) (St. Peterburg, Rusia, 1997)

[19] Hsu. C.S. And Dung M-T.Generating Finite-State Transducers for Semi-Structured Data Extraction from web. Information Systems 23, 8 (1998)

[20] Huck, G., Fankhauser, P., Aberer K., Nad Newhold E. J. Jedi: Extracting and Synthesizing information from the web. In proceedings of third IFCIS International Conference on Cooperative Information Systems (new York City, New York, 1998)

[21] Ion Muslea, Rise :Repositionary of Online Information Sources used in Information extraction task.

[22] Kushmerrick N., Wrapper Induction: Efficiency and Expressiveness. Artificial Intelligence Journal 118, 12 (2000).

[23] Laender A. H. F., Ribeiro Neto, B. A. And Da Silva, A. S. Dbye Data Extraction by Example. Data and Knowledge Engineering (2001)

[24] Laender A. H. F., Ribeiro Neto, B. A. , Da Silva A.S., And Silva, E. S. Representing Web Data as Complex Objects. In Electronic Commerce and Web Technologies .

[25] Liu, L., Pu., C And Han, W.Xwrap. An Xml Enabled Wrapper Construction System for web Information Sources. In proceeding of 16th IEEE International Conference on Data Engineering (SanDiego, California, 2000)

[26] Ludascher B., Himmeroder R, Lausen G. May, W., And Schlepphorst, C. Managing semistructured data with florid : A deductive object-oriented perspective. Information Systems 23, 8 (1998)

[27] Mecca G., Atzeni P., Masci A., Merialdo P., And Sindoni G. The ARANUES web base management system. SIGMOD Record 27,2 (1998)

[28] Muslea I., Extraction Patterns for Information Extraction tasks: A survey in proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction (Orlando, Florida, 1999)

[29] Muslea I., Minton S., And Knoblock C.A. Hierarchical Wrapper Induction for semistructured Information Sources. Autonomous Agents and Multiagents.

[30] Papakonstantinou Y., Garcia-Molina, H., And Widom J. Object Exchange Across Heterogeneous Information Sources. In proceeding of IEEE 11th International Conference on Data Engineering.(Taipai, Taiwan, 1995)

[31] Ribiero-Neto, B. A. Laender, A. H. F. And Da Silva. A. S. Extracting Semistructured Data Through Examples. in proceedings of Eight ACM International Conference on Information and Knowledge management. (Kansas City, Missouri, 1999)

[32] Sahuguet, A., And Azavant, F. Building Intelligent web applications using lightweight wrappers. Data and Knowledge Engineering 36, 3 (2001)

[33] Soderlan S. Learning Information Extraction Rules for semi-structured and Free Text. Machine Learning 34, 3 (1999)

[34] Teixeira, J. S. A Comparative study of Approaches for semostructured Data Extraction.Master's Thesis. Department of Computer Science, Federal University of Minas Gerais, Brazil, 2001

[35] World Wide Web Consortium. W3C. THE DOCUMENT OBJECT MODEL.

[36] M. Richardson And P. Domingos.Markov Logic Networks.Machine Learning, 62.

[37] Satyajeet Nimgaonkar and SuryaprakhDuppala, "A survey on web content mining and extraction of Structured and Semi structured data"

[38] H. Kautz, B. Selman, and Y. Jiang.A general stochastic approach to solving problems with hard and soft constraints. In The satiability problem: theory and applications. AMS, 1997.