ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2015

Comparative Study on Annotation Techniques

Poonam V. Wankhede [1], Sachin N. Deshmukh [2]

Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS), India[1]

Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS), India [2]

ABSTRACT: The internet give a great level of good knowledge which is usually formatted for its users, which make it troublesome to extract relevant data from various sources. The WWW (World Wide Web) plays a major role as all kinds of information repository and has been so far very successful in broadcasting information to humans. For the encoded data units to be machine processable which is indispensible for many applications, like deep web data collection and internet comparison shopping , they need to grouping and allot a meaningful labels. An automatic annotation approach, first align a data units on a result page into dissimilar groups, such that same group data have the same meaning or semantics. For each group annotate it from different feature and collective the different annotations to predict a final annotation label.

KEYWORDS: Data alignment, Data annotation, Web databases, Wrapper generation.

I. INTRODUCTION

Data mining is the computational process of discovering patterns in large datasets. Extract information from a dataset and convert it into an understandable structure for further use is the main goal of data mining. One technique of data mining is web mining, web mining discover patterns from the web. Web mining is divided into three different types - they are web usage mining, web content mining and web structure mining.

Web usage mining - Web usage mining is the application of data mining techniques to find outusage patterns from Web data, in order to recognize and improved serve the needs of Web-based applications. Web usage mining consists of three phases, the first is preprocessing, second is pattern discovery, and third is pattern analysis.

Web content mining - Web content mining is the mining techniques, which is extraction and integration of useful data, information and knowledge from Web page content.

Web structure mining - Web structure mining, one of three categories of web mining, it is a tool used to identify the relationship between Web pages linked by information or gave link connection direct. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. This connection allows a web search engine to drag data relating to a search query directly to the linking Web page from the Web site(location) in which the content rests upon. This completion takes place from side to side use of spiders scanning the Web sites, calling back the home page, then linking the information through reference links to bring forth the specific page containing the desired information.

Search engines are very important tools for people to get to the vast information on the World Wide Web. In recently people studies that web searching ; behind email is the second most popular activities on the internet ,a large portion of deep web is database based. This type of search engine is referred as web databases. Result pages returned from a web database has multiple search records. In each search result record contain multiple data units. There is a highly demand for collection of data from multiple web databases(WDB's). Large source of structure data is the world wide web. It is vast and fastly increasing recent of information. There are various kinds of things such as product, people, etc embed in both statically and non statically(dynamically) generated web pages. For manually annotate data it require large human effort which severely limit their scalability. Consider online shopping site for example :The eBay's database structure

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2015

is in unorganized manner, therefore it is a very time consuming process and for that automatic annotation approach proposed. Automatic annotation approach consist of three phases. The first is alignment phase, second is annotation phase and third is annotation wrapper generation phase.

Phase 1: Alignment phase identify all the data units in the SRR(Search Result Records) then organize them into different groups, with each group corresponds to same concepts. Same meaning of data can group it helps to identify the common patterns and the features of data.

Phase 2: Annotation phase introduce multiple basic annotators with each exploiting one type of feature. Annotator is used to produce a label for the data units within their group, and probability model is adopted to determine the best appropriate label for each group.

Phase 3: Annotator wrapper generation phase generate automatic annotation wrapper (rules), which execute annotation fast ,which is necessary for online applications.

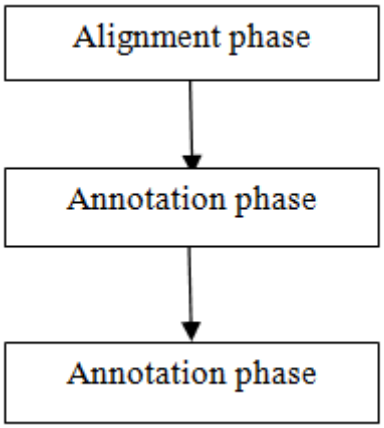

Fig 1. Three phase of annotation

II. TECHNIQUES

Automatic annotation approach improve the efficiency and reduce manual effort or human effort, most techniques are focus on automatic approaches in an alternative of manual or semiautomatic.

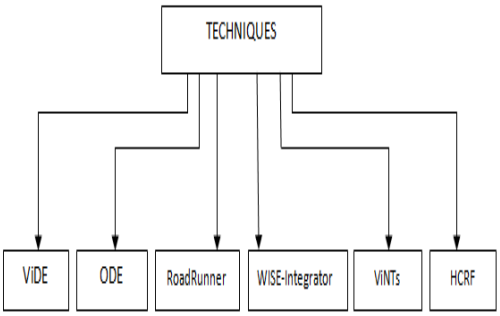

Fig 2. Recommendation Techniques

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2015

A. ViDE (Vision-based Data Extractor):
Extracting structured data from deep web pages is a difficult task due to the underlying complex structures of such pages. Some of the limitations are web page programming dependent or precisely HTML document and incapable of handling the ever increasing complexity of HTML source code. To overcome this problem Vision-based Data Extractor is proposed. ViDE [1] [2]is based on the visual feature, users can capture on the deep web pages while also utilizing some simple no visual information such as data type and frequent symbols to make the result more strong. In earlier times labeling is done manually, it is time consuming and errors can be occurred. After that semi automatic annotation came in to exist, in this approach there is no scalability. Then automatic annotation approach came in to exist. Some of the disadvantages of the Vision-based Data Extractorare,it can only process deep WebPages contain one data region, while there is number of multi-data region deep Webpage, which is a time consuming process.
Vision-based data extraction consists of two main components, Visionbased Data Record extractor (ViDRE) and Vision-based Data Item extractor (ViDIE).

B. ODE(Ontology-Assisted Data Extraction):
ODE is the Ontology-Assisted Data Extraction which automatically extracts the records from the HTML (Hyper Text Markup Language) files. Ontology-Assisted Data Extraction is accurate in identifying the query result [3], segmentation the query result into records and aligning and labeling the data values in the query result records. For many applications automatic record extraction is important like meta-querying, data warehousing etc. Data extraction is fully automatic and understanding of the query result page semantically. In Ontology Assisted Data Extraction, in semi automatic annotation there is no extra data are extracted the user can label only the data in which is interested.
Its drawbacks are time consuming and labor intensive, hence it is not applicable to large websites. To overcome the problem of semi automatic wrapper introduction some of the unsupervised method are being used such as Roadrunner [4], Omni [5] is fully automatically extracting the data from the query result page based on the tag structure that exist on HTML pages. To overcome these visual features it has been used for data extraction.

C. RoadRunner:
It is technique for extracting the data from HTML sites through the use of automatically generated wrappers. This technique is to automate the wrapper generation and the data extraction process to compare the HTML pages and introduce a wrapper based on their similarities and differences. Data is extracted by software modules called wrappers Manually coded wrappers is quite difficult, labor intensive task and difficult to maintain. The goals of fully automatic wrapper generation are,

- Works by using additional information.
- Assumption that the wrapper induction system has some priori knowledge.
- Generation of a wrapper by examining one HTML page at a time.

D. WISE-Integrator:
Searching is carried out either manually or semi automatically which is inefficient and difficult to maintain. It is a difficult task for users to access numerous Web sites individually to get the desired information. [6] H. He et.al it is a tool that performs automatic integration of Web Interfaces of Search Engines. It is used for identifying the matching attributes from different search interfaces for integration. WISE-extractor is capable of automatically grouping elements into logical attributes and deriving a rich set of meta-information for each attribute.

E. ViNTs (Visual information and Tag Structure):
A technique Visual information and Tag system [7] for automatically producing wrappers, used to extract search result record from dynamically generated result page. Automatic extraction of search result record is important for many applications. ViNTs utilizes both the visual content features on the result page as displayed on a browser and the HTML tag structure of the source file. Manually generating search result record wrappers is costly, time consuming and impractical for many application. Visual information And Tag structure based wrapper generator is a tool for automatically producing wrappers. In visual information and tag system, how to extract the dynamically generated search result pages returned by search engine. A result page contains multiple SRR's and some of the irrelevant

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2015

information to the users query. Accurate wrappers entirely based on the HTML tag structure. This method makes less sensitive to the misuse of the HTML tags.

F. HCRF(Hierarchical Conditional Random Field):
It is a Hierarchical Conditional Random Field [8] Existing approaches use decoupled strategies. That is attempting to data record detection and attributes labeling in two separate phases. Separately extracting data records and attributes is highly ineffective and proposes a probabilistic model to perform both processes simultaneously. Hierarchical Conditional Random Field can integrate all useful features by learning by their significance, and it can also incorporate hierarchical interaction. It is a template dependent. Expensive is the one of the main limitation.

III. ANNOTATION APPROACHES

There are three annotation approaches [9] for extracting information from web pages

- **A. Manual Approaches**
- **B. Semiautomatic Approaches**
- **C. Automatic Approaches**

A. **Manual Approaches:**
The initial approaches are the manual approaches in which languages were designed to assist programmer in constructing wrappers to identify and extract all the desired data items/fields. Labeling some sample pages, which is labor-intensive and more time consuming.

B. **Semiautomatic Approaches:**
Semiautomatic techniques can be classified into sequence-based and tree-based. The former, like as WIEN [10], Soft-Mealy [11], and Stalker [12], which represents documents as a series of tokens or characters and generates delimiter based extraction rules through a set of training examples. The latter, like as W4F [13] and XWrap [14], parses the document into a hierarchical tree (DOM tree), based on which they do extraction process. These approaches need manual efforts, like labeling some sample pages, which is human effort and time-consuming.

C. **Automatic Approaches:**
Reducing the manual effort and organized to improve the efficiency most new researches focus on automatic approaches in its place of manual and the other one is semiautomatic. Omni and RoadRunner is also one of the automatic approach[15]. The automatic data alignment method in[16] proposes a clustering approach to do alignment based on five features of data items, including font of text. Though this approach is mainly text based and tag structure based.

IV. COMPARATIVE ANALYSIS

A lot of system working on human interface [17] marks the preferred information on a sample pages and at the same time label these marked data [18] even the system can make a series of rules (wrapper) from the same source the same set of information are extracted on the web pages. These systems are referred as a wrapper induction system. These systems usually achieve high extraction accuracy because of the supervised training and learning process. But they suffer from poor scalability [19] and not suitable for applications [20] need to extract information from a large number of web sources.

Embley et al. [21] make use of ontologies together with several heuristics to automatically extract data in multi record documents and then label them. On the other hand, ontologies for different domains must be constructed manually.
Mukherjee et al. [22] proposed an annotation method but they are not fully automatic. Mukherjee develop the presentation styles and the spatial locality of semantically associated items, but for annotation its learning process is domain dependent and set of HTML documents are needs to be hand labeled. Therefore, these methods are not fully automatic.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2015

V. ADVANTAGES OF ANNOTATION SYSTEMS

Annotations have a lot of advantages such as follows.

- Static (Not moving) Type Checking- the compiler will check where the annotation is applicable.
- Clean Code- it's much easier to observe the Metadata defined in annotations.
- Readability
- Flexibility
- Length- XML-based configs are often very long. Annotations are much simpler to use.
- Active Process - Annotating is an active process. It's easy to become inattentive when simply highlighting.
- Ideas- Annotating helps you learn to pick out main ideas.
- Easy to Understand- Annotating helps user to understand.
- Easy to Remember- Annotating helps user remember.
- Additional Information- In web searching and dynamic link generation annotations provide the third party, relative metadata about the contents of web page to get additional information.

IV. CONCLUSION

We focus data annotation problem and proposed a multi annotator approach to automatically constructing an annotation wrapper for annotating the search result records retrieved from any given web database. Automatic annotation approach consists of six basic annotators and a probabilistic method to combine the basic annotators. Each of these annotators works one type of features for annotation.

REFERENCES

[1] W.Liu, X.Meng, and W.Meng," ViDE: A Vision-Based Approach for Deep Web Data Extraction," IEEE Trans .Knowledge and Data Eng., vol.22, no.3, pp. 447-460, Mar. 2010.
[2] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. World Wide Web (WWW), 2003.
[3] W.Su, J .Wang, and F.H. Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. Database Systems, vol. 34, no.2, article 12, June 2009.
[4] V. Crescenzi, G. Mecca, and P. Merialdo,"Road RUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc, Very Large Data Bases (VLDB) Conf., 2001.
[5] H.He, W.Meng, C.Yu, and Z. Wu," Automatic Integration of Web Search Interfaces with WISE-Integrator,"VLDB J., vol.13, no.3, pp.256-273, Sept.2004
[6] Yiyao Lu, Hai He, Hongkun Zhao, WeiyiMeng "Annotating Search Results from Web Databases "IEEE Transaction on Knowledge and Data Engineering, Vol. 25, No.3, March 2013
[7] H.Zhao, W. Meng, Z.Wu, V.Raghavan, and C.Yu,"Fully Automatic Wrapper Generation for Search Engines," Proc. Int'I Conf. World Wide Web, 2005.
[8] J. Zhu, Z. Nie, J. Wen, B.Zhang, and W-Y.Ma," Simultaneous Record Detection and Attribute Labeling in Web Data Extraction,"Proc. ACM SIGKDD Int'I Conf. Knowledge Discovery and Data Mining, 2006.
[9] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu, "Annotating Structured Data of the Deep Web," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), 2007.
[10] N. Kushmerick, "Wrapper Induction: Efficiency and Expressiveness," Artificial Intelligence, vol. 118, nos. 1/2, pp. 15-68, 2000.
[11] C.-N. Hsu and M.-T.Dung, "Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web," Information Systems, vol. 23, no. 8, pp. 521-538, 1998.
[12] I. Muslea, S. Minton, and C.A. Knoblock, "Hierarchical Wrapper Induction for Semi- Structured Information Sources," Autonomous Agents and Multi-Agent Systems, vol. 4, nos. 1/2, pp. 93-114, 2001.
[13] A. Sahuguet and F. Azavant, "Building Intelligent Web Applications Using Lightweight Wrappers," Data and Knowledge Eng., vol. 36, no. 3, pp. 283-316, 2001.
[14] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-sEnabled Wrapper Construction System for Web Information Sources," Proc. Int'l Conf. Data Eng. (ICDE), pp. 611-621, 2000.
[15] D. Buttler, L. Liu, and C. Pu, "A Fully Automated Object Extraction System for the World Wide Web," Proc. Int'l Conf. Distributed Computing Systems (ICDCS), pp. 361-370, 2001.
[16] Y. Lu, H. He, H. Zhao, W. Meng, and C.T. Yu, "Annotating Structured Data of the Deep Web," Proc. Int'l Conf. Data Eng. (ICDE), pp. 376-385, 2007.

ISSN(Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2015

[17] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997
[18] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources," Proc. IEEE 16th Int'l Conf. Data Eng. (ICDE), 2001.
[19] W. Meng, C. Yu, and K. Liu, "Building Efficient and Effective Metasearch Engines," ACM Computing Surveys, vol. 34, no. 1, pp. 48-89, 2002.
[20] Z. Wu et al., "Towards Automatic Incorporation of Search Engines into a Large-Scale Metasearch Engine," Proc. IEEE/WIC Int'l Conf. Web Intelligence (WI '03), 2003.
[21] D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y. Ng, and R. Smith, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31, no. 3, pp. 227-251, 1999.
[22] S. Mukherjee, I.V. Ramakrishnan, and A. Singh, "Bootstrapping Semantic Annotation for Content-Rich HTML Documents," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2005.