ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

Machine Learning Based Annotating Search Results from Web Databases

[1]P.Renukadevi,[2]K.Priyanka,[3]D.Shree Devi,[4]H.Abhijith,[5]T.Yogananth, [6]A.M.Ravishankkar, [7]Dr.S.Rajalakshmi

[1,2,3,4]Student, Department of Computer Science and Engineering, Jay Shriram Group of Instutions, Avinashipalayma, Tirupur, Tamilnadu, India

[5,6]Assistant Professor, Department of Computer Science and Engineering, Jay Shriram Group of Instutions, Avinashipalayma, Tirupur, Tamilnadu, India

[7]H.O.D., Department of Computer Science and Engineering, Jay Shriram Group of Instutions, Avinashipalayma, Tirupur, Tamilnadu, India

ABSTRACT: Deep web is a database based, i.e., for many search engines, data encoded in the returned result pages come from the underlying structured databases. Such type of search engines is often referred as Web databases (WDB). A typical result page returned from a WDB has multiple search result records (SRRs). Unfortunately, the semantic labels of data units are often not provided in result pages. Having semantic labels for data units is not only important for the above record linkage task, but also for storing collected SRRs into a database table. Early applications require tremendous human efforts to annotate data units manually, which severely limit their scalability. In this paper, we consider how to automatically assign labels to the data units within the SRRs returned from WDBs improve the results with new kernel function for improving the accuracy of the Support Vector Machines (SVMs) classification. The proposed kernel function is stated in general form and is called Gaussian Radial Basis Polynomials Function (GRPF) that combines both Gaussian Radial Basis Function (RBF) and Polynomial (POLY) kernels. We implement the proposed kernel with a number of parameters associated with the use of the SVM algorithm that can impact the results.

KEYWORDS: Data alignment, data annotation, web database, wrapper generation

I. INTRODUCTION

Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it. Knowledge discovery differs from traditional information retrieval from databases. In traditional DBMS, database records are returned in response to a query; while in knowledge discovery, what is retrieved is not explicit in the database. Rather, it is implicit patterns. The Process of discovering such patterns is termed as data mining. Due to enormous volumes of data, human analysts with no special tools can no longer find useful information. However, Data mining can automate the process of finding relationships and patterns in raw data and results can be utilized in an automated decision support system or assessed by a human analyst. That is why the data mining is very useful, especially in science and business areas which need to analyze large amounts of data to discover trends in it. The data mining would be one of the valuable assets, if we know how to reveal valuable knowledge that is hidden in the raw data. The data mining is a tool to extract diamonds of knowledge from the historical data and can also predict the outcomes of future situations.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

A. *Introduction to the project*

Data unit is a piece of text that semantically represents one concept of an entity. It corresponds to the value of a record under an attribute. It is different from a text node which refers to a sequence of text surrounded by a pair of HTML tags. Section 3.1 describes the relationships between text nodes and data units in detail. In this research, we perform data unit level annotation. There is a high demand for collecting data of interest from multiple WDBs. For example, once a book comparison shopping system collects multiple result records from different book sites, it needs to determine whether any two SRRs refer to the same book. The ISBNs can be compared to achieve this. If ISBNs are not available, their titles and authors could be compared. The system also needs to list the prices offered by each site. Thus, the system needs to know the semantic of each data unit. Unfortunately, the semantic labels of data units are often not provided in result pages. The semantic labels for the values of title, author, publisher, etc., are given. Having semantic labels for data units is not only important for the above record linkage task, but also for storing collected SRRs into a database table (e.g., Deep web crawlers) for later analysis. Early applications require tremendous human efforts to annotate data units manually, which severely limit their scalability. In this research we consider how to automatically assign labels to the data units within the SRRs returned from WDBs. The rules for all aligned groups, collectively, form the annotation wrapper for the corresponding WDB, which can be used to directly annotate the data retrieved from the same WDB in response to new queries without the need to perform the alignment and annotation phases again. As such, annotation wrappers can perform annotation quickly, which is essential for online applications.

B. *Motivation for work*

Consider the automatic data alignment problem in the annotation paper. Accurate alignment is critical to achieving holistic and accurate annotation. One method is a clustering based shifting method utilizing richer yet automatically obtainable features. This method is capable of handling a variety of relationships between HTML text nodes and data units, including one-to-one, one-to-many, many-to-one, and one-to-nothing. The experimental results show that the precision and recall of this method are both above 98 percent. Need to enhance our method to split composite text node when there are no explicit separators. So in this paper we using different machine learning techniques and using more sample pages from each training site to obtain the feature weights so that we can identify the best technique to the data alignment problem

II. LITERATURE SURVEY

A. *ViDE: A Vision-Based Approach for Deep Web Data Extraction*

Number of Web databases has reached 25 millions according to a recent survey . All the Web databases make up the deep Web (hidden Web or invisible Web). Often the retrieved information (query results) is enwrapped in Web pages in the form of data records. These special Web pages are generated dynamically and are hard to index by traditional crawler based search engines, such as Google and Yahoo. In this paper, we call this kind of special Web pages deep Web pages. Each data record on the deep Web pages corresponds to an object. Extracting structured data from deep Web pages is a challenging problem due to the underlying intricate structures of such pages. Until now, a large number of techniques have been proposed to address this problem, but all of them have inherent limitations because they are Web-page-programming-language dependent. As the popular two-dimensional media, the contents on Web pages are always displayed regularly for users to browse. This motivates us to seek a different way for deep Web data extraction to overcome the limitations of previous works by utilizing some interesting common visual features on the deep Web pages. In this paper, a novel vision-based approach that is Web-page programming- language-independent is proposed. This approach primarily utilizes the visual features on the deep Web pages to implement deep Web data extraction, including data record extraction and data item extraction.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

B. *On Deep Annotation*

Several approaches have been conceived (e.g. CREAM, MnM, or Mindswap) that deal with the manual and/or the semiautomatic creation of metadata from existing information. These approaches, however, as well as older ones that provide metadata, e.g. for search on digital libraries, build on the assumption that the information sources under consideration are static, e.g. given as static HTML pages or given as books in a library. Nowadays, however, a large percentage of Web pages are not static documents. On the contrary, the majority of Web pages are dynamic.2 For dynamic web pages (e.g. ones that are generated from the database that contains a catalogue of books) it does not seem to be useful to manually annotate every single page. Rather one wants to "annotate the database" in order to reuse it for one's own Semantic Web purposes. For this objective, approaches have been conceived that allow for the construction of wrappers by explicit definition of HTML or XML queries or by learning such definitions from examples. Thus, it has been possible to manually create metadata for a set of structurally similar Web pages. The wrapper approaches come with the advantage that they do not require cooperation by the owner of the database. However, their shortcoming is that the correct scraping of metadata is dependent to a large extent on data layout rather than on the structures underlying the data. While for many web sites, the assumption of non-cooperatively may remain valid, we assume that many web sites will in fact participate in the Semantic Web and will support the sharing of information. Such web sites may present their information as HTML pages for viewing by the user, but they may also be willing to describe the structure of their information on the very same web pages. Thus, they give their users the possibility to utilize 1. information proper,2. information structures and 3. information context. The success of the Semantic Web crucially depends on the easy creation, integration and use of semantic data. For this purpose, we consider an integration scenario that defies core assumptions of current metadata construction methods.In order to create metadata, the framework combines the presentation layer with the data description layer — in contrast to "conventional" annotation, which remains at the presentation layer. Therefore, we refer to the framework as deep annotation

C. *Automatic Annotation of Data Extracted from Large Web Sites*

Automatic systems leverage on the observation that data published in the pages of very large sites usually come from a back-end database and are embedded within a common HTML template. Therefore many pages share a common structure, and differences correspond to the data coming from the database. The wrapper generation process aims at inferring a description of the common template, which is then used to extract the embedded data values. These proposals reduce but do not eliminate the need for a human intervention. Since wrappers are built automatically, the values that they extract are anonymous and a human intervention is still required to associate a meaningful name to each data item. The automatic annotation of data extracted by automatically generated wrappers is a novel problem, and it represents a step towards the automatic extraction and manipulation of web data.

III. EXISTING SYSTEM

An increasing number of databases have become web accessible through HTML form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine process able, which is essential for many applications such as deep web data collection and Internet comparison shopping, they need to be extracted out and assigned meaningful labels. In this paper, we present an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantic. Then, for each group we annotate it from different aspects and aggregate the different annotations to predict a final annotation label for it. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

IV. PROPOSED SYSTEM

Kernel function called Gaussian Radial basis Polynomial Function (GRPF) is introduced that could improve the classification accuracy of Support Vector Machines (SVMs) for both linear and non-linear data sets. The aim is to train Support Vector Machines (SVMs) with different kernels compared with back-propagation learning algorithm in defining the class labels values for learning phase of the text. Moreover, we compare the proposed algorithm to algorithms based on both Gaussian and polynomial kernels by application to a variety of non-separable data with several attributes. Improves indicate that the proposed approach is highly effective. Proposed machine learning system used to annotate new result pages from the same web database with more accuracy than the existing learning algorithm.

V. SYSTEM ARCHITECTURE

It represents the Architecture of Annotating search results from Search result records.

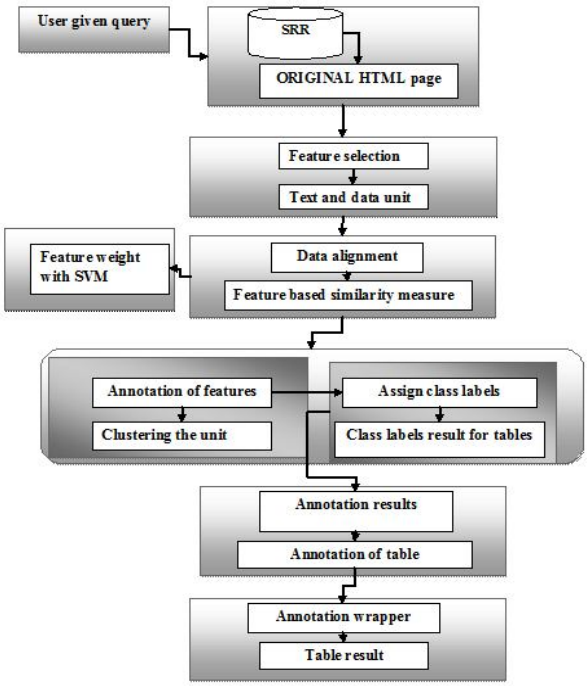

Fig.1. Architecture of Annotating search results from Search result records.

A. *Problem definition*

When the search result record pages are not available, their titles and authors could be compared. The system also needs to list the prices offered by each site. Thus, the system needs to know the semantic of each data unit. Unfortunately, the semantic labels of data units are often not provided in result pages. Early applications require tremendous human efforts to annotate data units manually, which severely limit their scalability. A large portion of the deep web is database based, i.e., for many search engines, data encoded in the returned result pages come from the underlying structured databases

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

B. *Support Vector Machine*

Support vector machine classification is choosing a suitable kernel of SVMs for a particular application, i.e. various applications need different kernels to get reliable classification results. It is well known that the two typical kernel functions often used in SVMs are the radial basis function kernel and polynomial kernel. More recent kernels are presented to handle high dimension data sets and are computationally efficient when handling non-separable data with multi attributes. However, it is difficult to find kernels that are able to achieve high classification accuracy for a diversity of data sets. In order to construct kernel functions from existing ones or by using some other simpler kernel functions as building blocks, the closure properties of kernel functions are essential.

For given non-separable data, in order to be linearly separable, a suitable kernel has to be chosen. Classical kernels, such as Gauss RBF and POLY functions, can be used to transfer non-separable data to separable, but their performance in terms of accuracy is dependent on the given data sets. The following POLY function performs well with nearly all data sets, except high dimension ones :

$$POLY\ x,z=(xTz+1)d$$

where d is the polynomial degree. The same performance is obtained with the Gauss RBF of the following form:

$$RBF\ x,z=exp\ \ (-\gamma|x-z|2)$$

where γ is appositive parameter controlling the radius. The Polynomial Radial basis Function (PRBF) as:

$$PRBF=((1+exp\omega)/v)d$$

where $\omega=x-z$ and $V=p*d$ is a prescribed parameter. Completely achieving a SVM with high accuracy classification therefore, requires specifying high quality kernel function,

Gauss RBF

Combine POLY, RBF, and PRBF into one kernel to become:

$$GRPFx,z=d+r.exp(-X-zr/(r.\sigma2))r+dd+1$$

$$\theta0=arg\ min\theta T(\alpha0,\theta)$$

where σ is a statistic distribution of the probability density function of the input data; and the values of r(r >1) and d can be obtained by optimizing the parameters using the training data. The proposed kernel has the advantages of generality. However, The existing kernels such as PRBF and proposed Gaussian and polynomials kernel function by setting d and r in different values. For example if d =0, we get Exponential Radial when r= 1 and Gaussian Radial for r= 2 and so on. Moreover various kernels can be obtained by optimizing the parameters using the training data .GRPF depends on two parameters d and r, encoded into a Vector $\theta = (d, r)$. We thus consider a class of decision functions parameterized by α,b,θ:

$$f\alpha,b,\theta x=sign(i=1l\alpha iyiGPRF\theta x,z+b)$$

and want to choose the values of the parameters α and θ such that w is maximized (maximum margin algorithm) and T, the model selection criterion, is minimized (best kernel parameters). More precisely, for θ fixed, we want to have

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

$\alpha 0 = \text{argmax}_w \alpha$ and choose $\theta 0$ such that $\theta 0 = \text{argmin}_\theta T(\alpha 0, \theta)$

When, θ is a one dimensional parameter, one typically tries a finite number of values and picks the one which gives the lowest value of the criterion T.When both T and the SVM solution are continuous with respect to h a better approach. They used an incremental optimization algorithm, one can train an SVMwith little effort when θ is changed by a small amount. However, as soon as h has more than one component computing $T(\alpha, \theta)$ for every possible value of h becomes intractable, and one rather looks for a way to optimize θ along a trajectory in the kernel parameter space. In this work, we use the gradient of a model selection criterion to optimize the model parameters. This can be achieved by the following iterative procedure:

1.Initialize θ to some value.

2.Using a standard SVM algorithm, find the maximum of the quadratic form w

$\alpha 0 = \text{argmax}_w \alpha$

3.Update the parameters h such that T is minimized. This is typically achieved by a gradient step

4.Go to step 2 or stop when the minimum of T is reached.

VI. CONCLUSION

In this research proposed a Kernel SVM based method for derive the weight values in the features that is text node and data unit nodes .If the feature weight values are derived automatically in the annotation phase after that performs the alignment phase using algorithm and then multi annotator approach to automatically constructing an annotation wrapper for annotating the search result records retrieved from any given web database. This approach consists of six basic annotators and a probabilistic method to combine the basic annotators. Each of these annotators exploits one type of features for annotation and our experimental results show that each of the annotators is useful and they together are capable of generating high-quality annotation. A special feature of our method is that, when annotating the results retrieved from a web database, it utilizes both the LIS of the web database and the IIS of multiple web databases in the same domain.

REFERENCES

1.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.
2. 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.
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. W. Meng, C. Yu, and K. Liu, "Building Efficient and Effective Metasearch Engines," ACM Computing Surveys, vol. 34, no. 1, pp. 48-89, 2002.
5. 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.
6. S. Handschuh and S. Staab, "Authoring and Annotation of Web Pages in CREAM," Proc. 11th Int'l Conf. World Wide Web (WWW),2003.
7. H. Zhao, W. Meng, and C. Yu, "Mining Templates form Search Result Records of Search Engines," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2007.
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'l Conf. Knowledge Discovery and Data Mining, 2006.