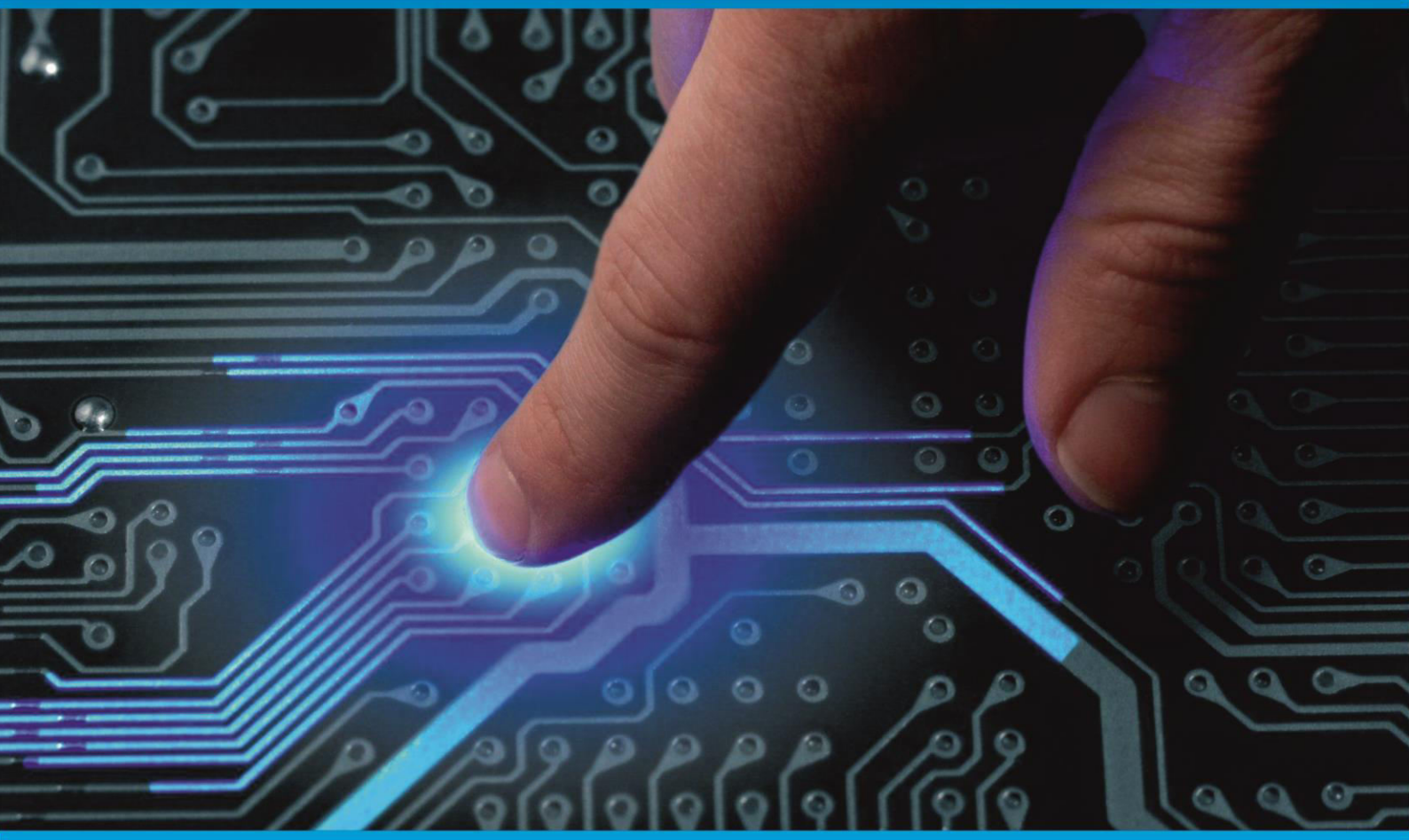
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 3, March 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903196|

An Enhanced Approach Based Automatic Annotation Results from Web Databases using ELM and Genetic Neuro Fuzzy Segmentation

P.JoySuganthy Bai[1],T.Mercy Grace[2]

Assistant Professor, Department of Computer Science and Engineering,GRACE College of Engineering, Thoothukudi Tamilnadu, India [1]

Assistant Professor, Department of Computer Science and Engineering,GRACE College of Engineering, Thoothukudi Tamilnadu, India [2]

ABSTRACT: The web has become the preferred medium for many database application. These application store information in huge databases that users can access, query and update through the web. The data units returned from the underlying database are usually encoded into result pages dynamically for human browsing. The data units are different from text node. The text node is surrounded by HTML tags. The data units are not used for application such as deep web data collection and internet comparison shopping. Hence annotation is done on the basis of data units. The data units are annotated by assigning meaningful labels to them. The automatic annotation method that first aligns the data units with their data unit similarity features.The Genetic Neuro Fuzzy segmentation algorithm is proposed for data alignment. Various types of annotators are used on the basis of data to be annotated. The proposed new machine learning algorithm called Extreme Learning Machine (ELM) which assign the input weight and calculate the weight. The ELM can be used to annotate new result pages from the same web database. This algorithm produce good performance and can learn thousand of times faster than other machine learning algorithm.

KEYWORDS: Data annotation, Data alignment, Extreme Learning Machine,

I. INTRODUCTION

Data mining, is the process of discovering interesting patterns and knowledge from huge amounts of data.The data sources can include databases, data warehouses, the web and other information repositories, or data that are streamed into the system dynamically.

The search engines are used to search the required information from the World wide web.The data returned from the Search engines are encoded into the returned result pages.These type of search engines are referred as the Web databases(WDB).The returned result page has several search result record (SRR) and each SRR has several entities.These entities are called as data unit and each dataunit is corresponds to the value of the record under an attribute.

The older application require human efforts for annotation,so the scalability is limited.The automatic annotation approach considers how to automatically assign labels to the data units.In the alignment stage ,the identified data units are organized into different groups.And each group having different concept of the same semantic.These grouping of same semantic for the data units are used to identify the similar patterns and features among the data units.In annotation stage multiple basic annotators are introduced and are used to produce label for the dataunits.

An annotation is a comment or explanation or presentation markup attached to text, image, or other data. An annotation stored in locally or in one or more annotation server. The local annotation can store annotation data in a local file system. Local annotation saved to the annotation directory. The remote annotation store annotation remotely on annotation server accessed through the web. Remote annotations are saved to the annotation post server.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903196|

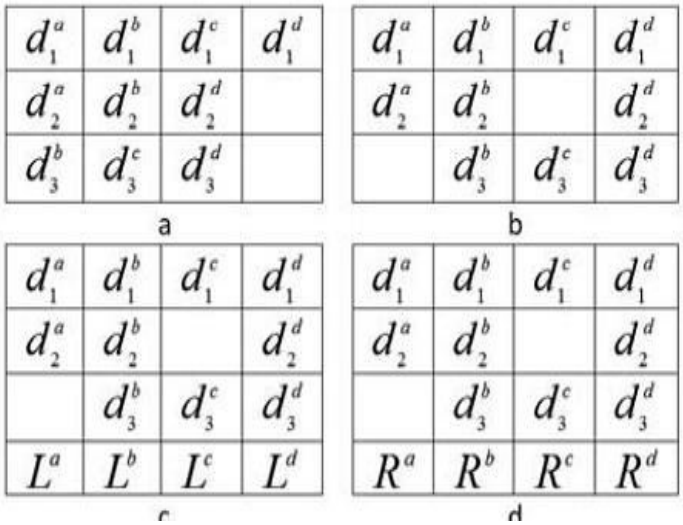


Fig.1 illustration of annotation process

Our automatic annotation process is illustrated in Fig.1.The d_i^j denote the data unit belonging to the ith SRR of concept j. The SRRs on a result page can be represented in a table format with each row representing an SRR. Each data unit aligned into different groups such that each groups have the same semantic. Label is assigned for each group using the basic annotators. An annotation wrapper for the search site is automatically constructed and can be is used to annotate the new web pages from the same WDB.

Our system has the following contributions

- Here we analyse the relationship between the text node and data units.
- Also we consider some important features such as datatype(DT), data content(DC),Presentation Style(PS) and adjacency information(AD).
- Here we use six basic annotators which results are combined to form a single label.
- Then new wrapper generation technique are introduced which are used to annotate new result pages for the same WDB.

II. RELATED WORK

N.Krushmerick, D.Weld and R.Doorenbos[5]proposed the wrapper induction system. A wrapper is a procedure for extracting tuples from a particular information source. Formally, a wrapper is a function from a page to the set of tuples it contains. Many systems rely on human users to mark the desired information on sample pages and label the marked data at the same time, and then the system induce a series of rules to extract the same set of information on web pages from the same source. Because of supervised training and learning process, these systems achieve high extraction accuracy but suffer from poor scalability.

Embley et al[4]proposed a conceptual-modeling approach to extract and structure data automatically. The approach is based on an ontology -a conceptual model instance - that describes the data of interest, including relationships, lexical form, and context keywords. By parsing the ontology, it automatically produce a database scheme and recognizers for constants and keywords, and then invoke routines to recognize and extract data from unstructured documents and structure it according to the generated database scheme. This technology achieves good recall and precision ratios for documents that are rich in recognizable constants and narrow in ontological breadth. However, ontologies for different domains must be constructed manually.

Arlotta et al[2] proposed annotation of data units with the closest labels on result pages. This method has limited applicability because many WDBs do not encode data units with their labels on result pages.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903196|

III. DATA ANNOTATION

A.Data Unit and Text Node Relationship

Text nodes are the visible elements on the webpage and data units are located in the text nodes.Text nodes are not always identical to data units. Annotation is at the data unit level need to identify data units from text nodes.A text node may contain to identify the following four types of relationships between data unit (U) and text node (T).

- One-to-One Relationship (denoted as T = U). In this type, each text node contains exactly one data unit.
- One-to-Many Relationship (denoted as T ⊐ U). In this type, multiple data units are encoded in one text node.
- Many-to-One Relationship (denoted as T ⊂ U). In this case, multiple text nodes together form a data unit.
- One-To-Nothing Relationship (denoted as T ≠ U). The text nodes belonging to this category are not part of any data unit

B. Data Unit and Text Features

These features are suitable to text nodes, including composite text nodes involving the same set of concepts, and template text nodes.

Data Content (DC)
The data units or text nodes with the similar concept often share certain keywords.

Presentation Style (PS)
This feature describes how a data unit is displayed on a webpage. It consists of style features such as font size color etc.

Data Type (DT)
Each data unit has its individual semantic type although it is just a text string in the HTML code such as Date time, string integer etc.

Tag Path (TP)
A tag path of a text node is a series of tags traversing from the root of the SRR to the corresponding node in the tag tree.

C.Data Unit Similarity

Two data units belong to the same concept is determined by how similar they are based on the Features.The similarity between two data units (or two text nodes) is a weighted sum of the similarities of the five features.

$$Sim(d_1,d_2)=W_1*SimC(d_1,d_2)+W_2*SimP(d_1,d_2)+W_3*SimD(d_1,d_2)+W_4*SimT(d_1,d_2)+ W_5*Sim A(d_1,d_2) \quad (1)$$

D. Basic Annotators

Table Annotator (TA)
First, it identifies all the column headers of the table.
Second, for each SRR, it takes a data unit in a cell and selects the column header whose area (determined by coordinates) has the maximum vertical overlap (i.e., based on the x-axis) with the cell. This data unit is then assigned with this column header and labeled by the header text A .

Query-Based Annotator (QA)
Given a query with a set of query terms submitted against an attribute A on the local search interface. First get the group that has the largest total occurrences of these query terms and then assign gn(A) as the label to the group.

Schema Value Annotator (SA)
The schema value annotator first identifies the attribute that has the highest matching score among all attributes and then uses gn (Aj) to annotate the group Gi.

Frequency-Based Annotator (FA)
The frequency-based annotator intends to find common preceding units shared by all the data units of the group Gi.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903196|

E. Combining Annotators

Based on this characteristic, we employ a simple probabilistic method to combine different annotators. For a given annotator L, let P (L) be the probability that L is correct in identifying a correct label for a group of data units when L is applicable.P (L) is essentially the success rate of L. Specifically suppose L is applicable to N cases and among these cases M are annotated correctly, then P (L) =M/N.

IV. PROPOSED SYSTEM

A.Extreme Learning Machine

The Machine Learning is the science of getting computers to act without being explicitly programmed. It is the sub field of Computer science and statistics that deals with the construction and study of systems that can learn from data rather than follow any explicitly programmed instructions.

Neural Networks have been extensively used in many fields due to their ability to approximate complex nonlinear mappings directly from the input sample; and to provide models for a large class of natural and artificial phenomena that are difficult to handle using classical parametric methods. There are many algorithm for training Neural Network like Back propagation, Support Vector Machine (SVM), Hidden Markov Model (HMM) etc. One of the disadvantages of the Neural Network is the learning time.

The proposed new learning algorithm called Extreme Learning Machine (ELM) which overcomes the problems caused by gradient descent based algorithms such as Back propagation.

- ELM can notably reduce the amount of time needed to train a Neural Network.
- The learning speed of ELM is very fast.
- The ELM has improved generalization performance than the gradient-based learning such as back-propagation
- It requires less human being interventions and can run thousands times faster than those conventional methods
- The ELM suitable for many nonlinear activation function and kernel functions.

The traditional classic gradient-based learning algorithms may face several issues like local minima, inappropriate learning rate and over fitting, etc. In order to avoid these issues, some methods such as weight decay and early stopping methods may need to be used often in these classical learning algorithms. The ELM tends to reach the resolution straightforward without such trivial issues. The ELM learning algorithm seems to be much simpler than most learning algorithms for feedforward neural networks

B.ELM Algorithm

- Given training set with activation function and hidden number of nodes.
- Select the activation Function.
- Assign input weight and bias.
- Calculate the output matrix of hidden layers.
- Calculate the output weight of the hidden layers.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903196|

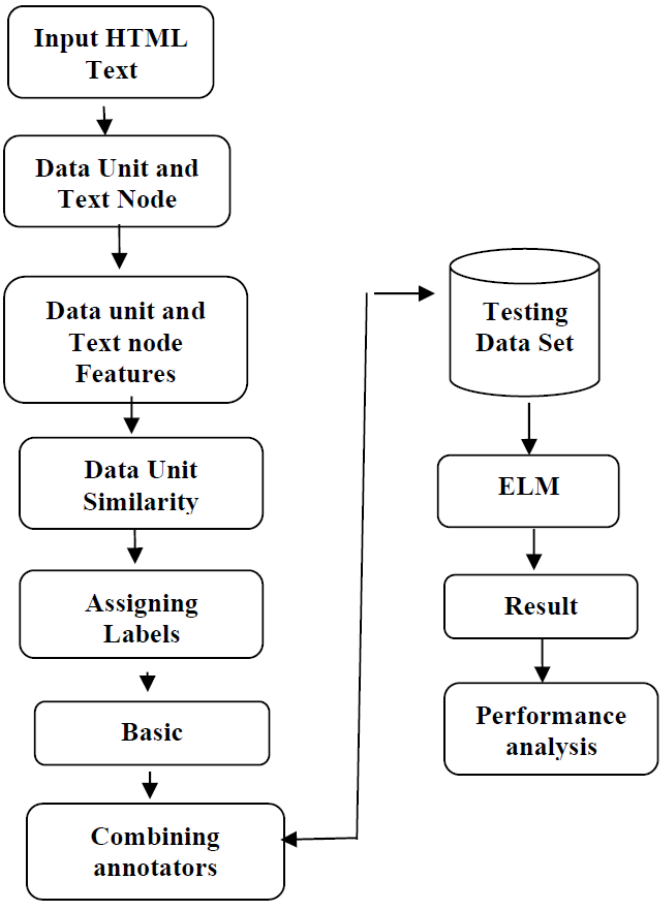


FIG.2. SYSTEM ARCHITECHTURE

C .***Genetic Neuro Fuzzy Segmentation Algorithm:***

Genetic Algorithm based neuro fuzzy system has following step:

1. Get input
2. Extract features and mark the features as feature 1and feature 2.
3. Parameters are set for Genetic Algorithm.
4. Initial Subset of 7 features is selected randomly from all possible solution sub space. These 7 features are represented by binary string (11010110101000000000), where 1 shows presence of feature and 0 absence of feature i.e. Feature no1, 2, 4, 5,6,7,9 are selected initially
5. Fitness Function: Fitness function decides the success of Genetic Algorithm. For proposed research work Fitness function is determined by neuro fuzzy

$$\text{Fitness}(f) = \sum_{i=1}^{\infty 20} \frac{TP}{TP+FN} + \frac{TN}{TN+FP} \frac{TP+TN}{TP+TN+FP+FN} \ldots \quad (2)$$

6. Fitness max represents maximum threshold value for a feature subset.
   Best_Feature_ Subset =Fitness(f)-Fitnessmax
   If the difference between two is 0 to 15 then that particular feature subset is considered as best feature subset.
   Otherwise mutation and crossover is done to generate new population.
7. Go to step 5 again

V. EXPRIMENTATION AND RESULT ANALYSIS

In this section, we discuss the experimentation and result analysis of the ELM algorithm. The result is embodied with performance evaluation of precision and recall technique. Precision and recall is calculated for performance of

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903196|

alignment and performance of annotation.Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is generally expressed as a percentage. The precision value only shows the total no of retrieved books from the website based on given topic. It computes only the highest number of books found in the different web pages and the percentage of overall book data found in all websites.

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is generally expressed as a percentage. Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.Recall represents only the highest number of relevant and repeated books found in the different web pages and the percentage of overall book data found in all websites. It provides only the total no. of percentage result.Fig.3. shows the annotation of different basic annotators.Fig.4.illustrates the evaluation of performance.

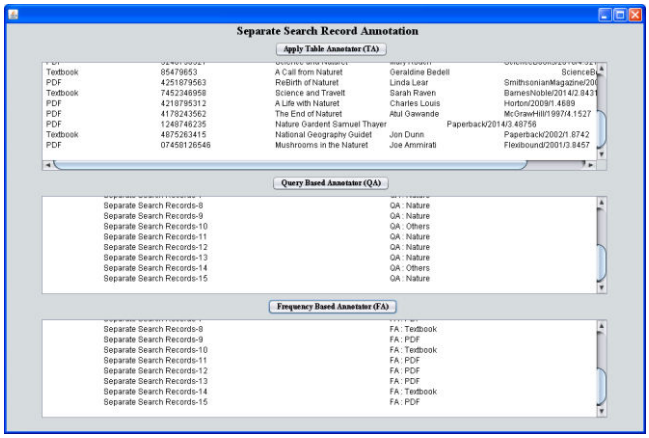


Fig.3 Annotation using basic annotators

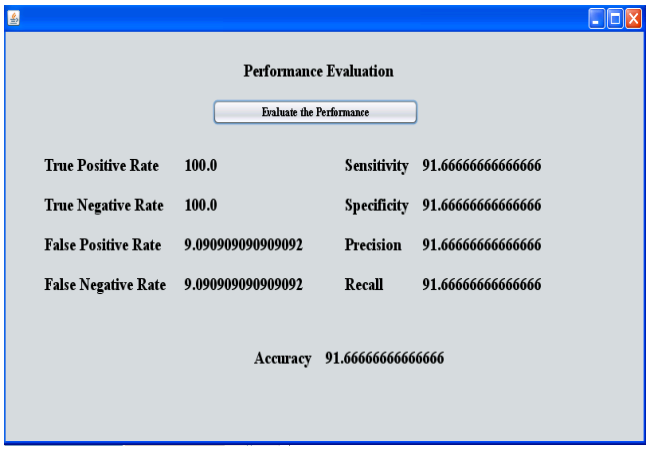


Fig.4 Performance Evaluation

VI. CONCLUSIONS

The automatic annotation approach considers several types of data unit and text node features and makes annotation scalable and automatic. The approach consists of six basic annotators and a probabilistic method to combine the basic annotators. Each of these annotators develops one type of features for annotation. A new Extreme learning machine algorithm for data annotation in the web database would be proposed. The proposed technique would be executed with the expected results by using knowledge database as a database and our experimental results show that each of the annotators is useful and they together are capable of generating high quality annotation.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 3, March 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0903196|

REFERENCES

[1] Arvind Arasu and Hector Garcia-Molina "Extracting Structured Data from Web Pages" ,Proc.SIGMOD Int'l Conf.Management of Data, 2003.

[2] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large Web Sites,"Proc.SixthInt'l Workshop the Web and Databases (WebDB),2003

[3] H. Elmeleegy, J. Madhavan, and A. Halevy, "Harvesting Relational Tables from Lists on the Web," Proc. Very Large Databases (VLDB) Conf.,2009

[4] 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

[5] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI),1997

[6] 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

[7] Nilesh Dalvi, Ravi Kumar and Mohamed Soliman "Automatic Wrappers for Large Scale Web Extraction" Proceedings of the VLDB Endowment, Vol. 4, No. 4 Copyright 2011.

[8] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases,"Proc. 12th Int'l Conf. World Wide Web (WWW),2003

[9] M.Yazhmozhi, M. Lavanya, Dr. N. Rajkumar "Annotating Multiple Web Databases Using SVM" International Journal of Innovative Research in Computer and Communication Engineering Vol.2, Special Issue 1, March 2014.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details