ISSN 2320 - 2602
Volume 3, No.3, March 2014
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2014/ijacst02332014.pdf

Semantic Similarity Based Data Alignment and Best Feature Extraction using PSO for Annotating Search Results from Web Databases

T.Seeniselvi[1], N.Thangamani.[2]
[1]Associate Professor, Hindusthan college of Arts and Science, India,seeniselvi@yahoo.co.uk
[2]Research Scholar, Hindusthan college of Arts and Science, India,thangamani360@gmail.com

ABSTRACT

Due to the development of search engines databases through web reachable all the way through HTML based search boundary in now day's analysis of data in deep manner from database or web search engines also important to return exact information in search result web pages. In generally the data units received from web accessible search engine databases are frequently prearranged into the result pages energetically for individual browsing. In this paper, consideration of automatic data assignment for SRRs pages returned from original web search engine databases. To conquer these problems proposed an automatic semantic annotation approach through semantic similarity measure for data units and text unit's results from features for Search results records. The features of data and text units are obtained from Particle Swarm Optimization (PSO) methods. From search results records important feature are extracted and then semantic similarity based measurement are measures are performed to each and every data, text unit nodes. Ontology based system measures semantic similarity between terms in the pages and then aligns the data units in efficient manner. In this work we proficiently analysis the data and most excellent alignment of SRR records. To annotation of new search result from web search engines for various domains in databases we use annotation wrapper. Our experimentation results are estimated based on the parameters like precision and recall for various topics.

Keywords: Annotating search results, Semantic similarity, Feature extraction, Particle Swarm Optimization, web database, wrapper generation.

1. INTRODUCTION

Due to the development of the search engines in now days analysis of data in deep manner from database or web search engines also important to return exact information in search result web pages. Since the entire deep web data are encoded in HTML format in search engines. These types of search engines maintained by system are known as Web databases. Several depth analysis methods have been performed in earlier years to extract information from encoded format pages, but still the search results of web deep analysis are not correctly aligned in efficient manner. Normal result page from web search engines includes

Numeral of search result records (SRR) pages and each one of the SRRs related to individual concepts. For example consider reliance mart online web page Encoded HTML page it consists of information about products .Usually each SRR consists of numerous data units like product name, brands, rate, company, etc. Normally, not every one of the data units in the SRRs are not encoded in meaningful manner or semantic manners. To conquer this problem in this paper proposed an efficient algorithm to automatically interpret the data units in the SRRs pages through Web databases. But these results not investigate SRRs search sites with the intention of enclose Web services interfaces, since exact semantic meaningful of data units labels are most accurately described in WSDL.

For examples no semantic labels for the principles of numerous data units like product name, brands, rate, company, are specified. Since semantic labels of each and every data units from SRR not simply significant for record association task, although it is important for accumulate gather SRRs into a table format for data analysis. Early applications necessitate incredible human efforts to interpret data units physically, which rigorously maximum value their scalability. In this paper, consideration of automatic data assignment for SRRs pages returned from original web search engine databases .Number of approaches have been used in literature to mine SRRs in well organized manner . The first move toward is the physical approach. Through examine a Web page and its basis code, the programmer discovers various examples from the page and after that writes a program to classify and extract every data units efficiently. This approach is not enough to extract data units from large number of pages. Other approaches the entire have various amount of computerization. The majority of existing approaches [1,2] simply allocate labels to everyone HTML text node; systematically evaluate the associations among text nodes and data units.

In this paper present an efficient automatic semantic annotation of label units in semantic manner and extract features from SRRs features such as text and data unit feature using Particle Swarm Optimization (PSO) methods. From search results records essential feature are extracted and then semantic relationship based measurement are achieve to each one and all data, text unit .The propose a clustering-based shifting procedure that group various results into single line consequently with the intention of the data units within the similar group have the equivalent semantic. Grouping of similar data unit with same meaning can help categorize the frequent patterns and features amongst these data units.

2. BACKGROUND STUDY

Several systems [3], [4] rely on human user to mark the preferred data on model page as well as label the noticeable information by the similar time, furthermore followed by the system can induce a sequence of rules (wrapper) to remove the similar set of information on WebPages since the similar source. These types of systems are frequently used as wrapper simulation system. Use of this system supervised instruction as well as observes process can be able to reach high extraction accurateness. Since, they are not suitable for applications because poor scalability [5], [6] those require extracting information as of huge number of network sources. A wrapping is a program so as to extract information since Web site or Web page furthermore put them into the database [7].Here use two main methods to wrapper creation.

The initial methods is known as wrapper induction methods which performs learning based on supervised learning methods to study data extraction units from rules those are physically labeled as positive and negative examples. Manual labeling of data is, conversely, work exhaustive and time consuming. Furthermore, for dissimilar sites, the labor-intensive labeling procedure needs to be frequent since they follow dissimilar pattern. Illustration wrapper induction scheme include WIEN [8, 9], WL2 [7, 10]], and the rest .Our procedure necessitate no character classification. It extracts data records in a SRRs pages and mine data from the records mechanically. The subsequent move toward is automatic extraction. In [11] proposition a small number of further heuristics to achieve the task not including by means of domain ontology. Conversely, [12] demonstrate with the intention of these methods construct concentrated outcome. In adding together, these methods do not mine information beginning data records.

Wrapper induction [13-14] is a semantic automation method that extracts data automatically from Web pages, but their extorted data is not explained. Though, ontologies for special domains should be assembling physically through a proficient [15] make use of the presentation styles and the spatial position of semantically associated units, however its learning procedure for explanation is domain-dependent. We are responsive of simply two recent works [16, 17] with the intention of intend at automatically assigning important labels to SRRs not including human relations and domain drawback. Due to the large number of web pages in the search engine results, it is not to form data table from encoded data in database.

3. DATA ALIGNMENT AND FEATURE EXTRACTION

In this paper proposed work first analysis deep web data based on the pages encoded in HTML format. To analysis data and text units important features are need to extracted and measure semantic similarity among those features ,aligned in proper manners into table .In this work feature of data text and units are extracted using Particle swarm optimization (PSO) algorithm .Similarity among those features are obtained through ontology algorithm .Before Extraction and measuring semantic similarity of the data and text unit we need many types of relationship among one page to another page are estimated using following relationship

One-to-One Relationship, every text node include accurately one data unit. This is the majority regularly see case. The every text in the encoded page enclosed by the pair of tags <A> and </A>.

One-to-Many Relationship every text node include multiple data units. Because the text of such category of nodes can be measured as a work of art of the texts of numerous data units, identify it a combination text node. This examination is suitable in universal since SRRs are creating by pattern programs.

Many-to-One Relationship numerous text nodes simultaneously type a data unit. It is a universal follow that webpage stylish use individual HTML tags to decorate confident information. For the principle of extraction and annotation, necessitate identifying and removing these tags within SRRs consequently that the completeness of every divide data unit can be returned.

One-To-Nothing Relationship every text unit nodes in the example belong to the category of text unit only not data unit within SRRs. It makes use of a frequency-based annotator to discover template text nodes.

3.1. FEATURE EXTRACTION USING PSO

Particle swarm optimization is one of most important optimization algorithm to solve many of the real time application problems based on the development and ability of particles. It uses particles to representation of HTML data unit pages as input that moving from one particle HTML pages to another HTML pages from SRRs .The major importance of the this process is to extract important features such as Data Content (DC), Presentation Style (PS), Data Type (DT), Tag Path (TP), and Adjacency (AD) from HTML encoded pages for web search engine. It used to extract feature efficiently from SRRs pages and improves web search return results. The important features are investigated by each particles in the pages is carry out through formulation of every particles as pages with known speed and location. Every SRRS extract the best features and remove the irrelevant features, moreover update position of current location SRRs pages to next SRRs pages in the investigation, at the same time reorganized by everyone and all step. Moreover, every page extracted feature from SRRs is kept in memory, detection the best features position of search space in SRRs it has progressively more visited. Thus, its grouping SRRs is a combined speeding up towards the furthermost features extraction of a topological neighborhood. PSO based feature extraction from SRRs is estimated and examination in excess of SRRs as well as global features in the SRRs are extracted efficiently.

Proposed PSO feature extracted from SRRs are organized according to web search results from, they often to find the best feature such as Data Content (DC), Presentation Style (PS), Data Type (DT), Tag Path (TP), and Adjacency (AD) from HTML encoded pages. After extraction of best features from SRRs, and then studies their associated to SRRs with similar concepts and generate suitable features with various type of concepts of j^{th} SRRs in i's SRRs simultaneously related to the same concept that are asked by user at the web search engine. Each SRRs in the particles are compared to another SRRs with associated to various concepts by the best efficient important features are extracted by any member of its current SRRs features p_i The vector p_i for that best local features for SRRs, which we indicate initially as global best extracted features for SRRs. Initialize the particle's location well-known best extracted features that are associated to the search engine result for user to its initial SRRs position: $p_i \leftarrow x_i$. then likewise update current best extracted features for SRRs and their velocity of best important features are extracted to each SRRs. Proposed algorithm repeats these above mentioned steps until all the best features are extracted from encoded HTML pages .Finally find best important features through global and local best features work based on PSO.

Particle swarm optimization (PSO) with local knowledge phase

1. Initialize a population as number of pages encoded SRRs is considered as particles with random location and rapidity on D dimensions in the search space of best local pages features extraction.
2. loop
3. For each pages from SRRs considered as particle, evaluate the desired optimization fitness function in D variables.
4. Compare particle's that is user SRRs with its $pbestloc_i$ to extract features . If current value is better than $pbestloc_i$, then set $pbestloc_i$ equal to the current value of the best features from SRRs
5. Initialize the particle's that is SRRs pages best known position to its initial position: $p_i \leftarrow x_i$
6. Identify the particles (SRRs) in the neighborhood with the best features extraction and assign its index to the variable g.
   6.1. If $(f(p_i) < f(g))$ update the swarm's best known position: $g \leftarrow p_i$
7. Initialize the particle's velocity: $v_i \sim U(-|b_{upv}-b_{lov}|, |b_{upv}-b_{lov}|)$
8. Change the velocity and position of the particle SRRs according to the following equation
9. Until a termination criterion is met ,repeat:
   9.1. For each particle (SRRs) (i = 1, ..., S)
   9.2. do
   9.3. For each dimension d = 1, ..., n
   9.4. do
   9.5. Pick random numbers $r_p, r_g \sim U(0,1)$
   9.6. Update the particle's velocity $v_{i,d} \leftarrow \omega\, v_{i,d} + \varphi p_{rp}\,(p_{i,d}-x_{i,d}) + \varphi grg\,(g_d-x_{i,d})$
   9.7. Update the particle's position with according to best features from SRR: $x_i \leftarrow x_i + v_i$
   9.8. If $(f(x_i) < f(p_i))$ do:
   9.9. Update the particle's best features extraction for SRRs position: $p_i \leftarrow x_i$
   9.10. If $(f(p_i) < f(g))$ update the particles best features known position: $g \leftarrow p_i$
   9.11. Now g holds the best found solution.
10. end loop.

The important features are mentioned below which features are only extracted from PSO algorithm .After relationship is identified between attributes and data unit then it is important to extract various features in the encoded page .So we use particle swarm optimization algorithm to extract features from SRRs pages. The most important features in the pages are Data Content (DC), Presentation Style (PS), Data Type (DT), Tag Path (TP), and Adjacency (AD).

Unit (DU) is defined as data which is similar to equal concepts with keywords; it is important feature to extract search results efficiently

Presentation Style (PS) is defined as different styles supported by web pages that are font style, size, color, text adornment, etc., and whether it is italic or bold. When the data unit belongs to same concept for different SRRs are supported by the same style.

Type (DT) is defined as type or category of data unit belongs to same concept is found. The subsequent essential data types are presently well thought-out in our move towards day, moment in time, exchange, numeral, Decimal, Percentage, character, and String.

Tag Path (TP) is defined as series of nodes present in the text node from HTML and navigates beginning the derivation of the SRR to the resultant node in the tag hierarchy.

Adjacency (AD) is defined as the results are found by keywords to search similar concepts and find from different SRRs.

3.2. SEMANTIC SIMILARITY MEASUREMENT USING ONTOLOGY

In this paper to differentiate various data units and identify similar data unit concepts through roughest theory based classification which identifies similarity among data unit's results from above similarity measures formation which partitions the creation into sets of related data units called elementary sets [17]. The elementary sets of the data units in the encoded file format can be used to create much information on the data and text units along with similarity functions use of similarity among various representations go ahead to information granulation. It assumes that every data units are encoded in HTML format of the creation is connected through a definite quantity of data and text units information, characterize by a number of attributes which communicate the descriptions of data units, text units. The entire description how to use data and text unit description for data alignment problems with OARS [18]. The concept of data units and text units in rough sets is demoralized in OARS to compact with doubts throughout the map procedure of ontology alignment when the results of data and text unit concepts are exactly matched between different data user with same. Using the similarity of data units and text unit's nodes in the pages considered as the attributes of elements for better grouping of similar concepts which is additional second-hand to conclude the similarities among the data units based on their characteristic standards. Algorithm 2 shows the pseudocode of the data alignment algorithm for data and text units. Line 1 is used to assign the six different similarity measures among data and text unit features which are extracted from PSO. Lines 2–6 are used to select the different attributes in SRRs for alignment based on the accuracy results. Lines 7–10 are used to allocate a self-confidence amount to the mapped data and text unit nodes from features using PSO. Let $[x]_F$ denote a set of data and text units amongst them with observe to known matching factors similarity results from equation (1) to equation (7).

Let

U Be the set of unmapped data units in the ontology $U = \{d_1, d_2, ., d_n\}$

F be the set of matched factors of each data units in the results $F = \{f_1, , \dots, f_n\}$

X Be subset of U

The accuracy result of data alignment units of the input feature extraction results of the set X can be computed using

$$\alpha_F(X) = \frac{|(X)|}{|F(X)|}$$

f_1 Represents the value of $Sim_{text}(d_i, d_i')$ as defined in(1)

f_2 Represents the value of $Sim_{datalines}(d_i, d_i')$and $Sim_{hsp}(d_i, d_i')$ as defined in(2) and (3) respectively

f_3 Represents the value $Sim_{pr}(d_i, d_i')$ & $Sim_{strc}(d_i, d_i')$ as defined in (4&5).

f_4 represents the value of $Sim_p(d_1, d_2)$ as defined in (6)

f_5 represents the value of $Sim_T(d_1 \& d_2)$ as defined in (7)

f_6 represents the value of $Sim_A(d_1, d_2)$ as defined in (8)

3.2.1. TEXT UNIT SIMILARITY

Measuring the semantic similarity among text nodes between two nodes with same concept remove the present text unit and group them into same concept. The differences of two similar two text unit nodes is defined as below used in Smoa with the lengths of without comparison strings.

$$Sim_{text}(d_i, d_i') = Smoa(d_i, d_i') \rightarrow (1)$$

3.2.2. DATA UNIT SIMILARITY

The data unit similarity of two data units for different text unit nodes are measured based on following conditions:

- $Sim_{datalines}(d_i, d_i')$ be the linguistic similarity among data units in the pages
- Σ be the external resource (wordnet)
- $s(d_i)$ be the set of data units with similar synonyms
- $h(d_i)$ be the set of hyponyms for data units and hypernyms for data units
- $t(d_i)$ be the set of antonyms that related to data units d_i'

The similarity of data units is measured as,

$$Sim_{datalines}(d_i, d_i') = \begin{cases} 1 & if\ d_i' \in s(d_i) \\ 0.5 & if\ d_i' \in h(d_i) \\ 0 & if\ d_i' \in t(d_i) \end{cases} \rightarrow (2)$$

3.2.3. DATA CONTENT SIMILARITY

The data content similarity is measured based on the following conditions with same concept that are encoded in HTML page, it need to satisfy the following conditions. $Sim_{hsp}(d_i, d_i')$ Be structural similarity between different data unitsd_i, d_i'.

$K_{sup}(d_i)$ Be the set of super classes of different data with same concept d_i

$K_{sup}(d_i')$ Be the set of super classes of different data with same concept d_i'

$|K_{sup}(d_i)|$ Be the set of super classes of cardinality different data with same concept $K_{sup}(d_i)$

$|K_{sup}(d_i^{'})|$ Be the set of super classes of cardinality different data with same concept $K_{sup}(d_i^{'})$

$$Sim_{hsp}(d_i, d_i^{'}) = \frac{1}{2}\left(\frac{|(K_{sup}(d_i) \cap K_{sup}(d_i^{'}))|}{|K_{sup}(d_i)}\right) + \left(\frac{|(K_{sup}(d_i) \cap K_{sup}(d_i^{'}))|}{|K_{sup}(d_i^{'})}\right) \rightarrow (3)$$

The similarity between the data unit's properties is also plays major important to find exact concept related similarity .Let us consider the data units characteristics to measure similarity. $Sim_{pr}(d_i, d_i^{'})$ Represents the similarity among data units with their properties

$pr\ (d_i)$ Be the set of super classes of different data properties with same concept d_i

$pr(d_i^{'})$ Be the set of super classes of different data properties with same concept $d_i^{'}$

$|pr(d_i)|$ Be the set of super classes of cardinality different data properties with same concept $pr(d_i)$

$|pr(d_i^{'})|$ Be the set of super classes of cardinality different data properties with same concept $pr(d_i^{'})$

$$Sim_{pr}(d_i, d_i^{'}) = \frac{1}{2}\left(\frac{|(pr(d_i) \cap pr(d_i^{'}))|}{|pr(d_i)}\right) + \left(\frac{|(pr(d_i) \cap pr(d_i^{'}))|}{|pr(d_i^{'})}\right) \rightarrow (4)$$

Finally combine both of this similarity into one

$$Sim_{strc}(d_i, d_i^{'}) = \frac{1}{3}\left(Sim_{hsp}(d_i, d_i^{'}) + Sim_{hsp}(d_i, d_i^{'}) + Sim_{pr}(d_i, d_i^{'})\right) \rightarrow (5)$$

3.2.4. PRESENTATION STYLE

Presentation style of the data units must find the different styles of units between data units d_1 & d_2

$$Sim_p(d_1, d_2) = \sum_{i=1}^{6} Ds_i / 6 \rightarrow (6)$$

Where Ds_i is the score of the i^{th} style data type defined by $Ds_i = 1$ if $D_d^1 = D_d^2$ and $Ds_i = 0$ otherwise, and D_d^I is the i^{th} style of data unit d.

3.2.5. TAG PATH SIMILARITY

Tag path similarity is defined as distance between two similar text units the tag tree. Let p_1 &p_2 be the tag paths of d_1 &d_2, correspondingly, and $PLen(p)$ indicate the numeral of tags in tag path p relationship among d_1 &d_2 is

$$Sim_T(d_1\ \&d_2) = 1 - \frac{EDT(p_1, p_2)}{Plen(p_1) + Plen(p_2)} \rightarrow (7)$$

3.2.6. ADJACENCY

The adjacency relationship among two data units d_1 &$d_2$is the normal of the relationship among ${d_1}^p$ &${d_2}^p$ and the relationship among${d_1}^s$ &${d_2}^s$ that is

$$Sim_A(d_1, d_2) = \frac{Sim^{'}(d_1^p, d_2^p) + Sim^{'}(d_1^s, d_2^s)}{2} \rightarrow (8)$$

Algorithm 2: Data alignment for data and text units

Input: $D = \{d_1, d_2, d_3, \dots, d_m\}$, a set of unmapped data units from the source ontology; $D^{'} = \{d_1^{'}, d_2^{'}, d_3^{'}, \dots, d_m^{'}\}$, a set of unmapped data units from target ontology $F_1 = F$, $F = \{f_1, f_2, . , . f_n\}$ a set of matching data units ,n=6

Output: Aligned units (d_i, d_j, c) where c is the best confidence degree

1. For k=1 to 6;
2. For i=1 to m;
3. For j=1 to n;
4. Compute $\alpha_f = 1$ THEN
5. ALIGN $(d_i, d_J{'})$
6. IF $F_k = F_I$ then
7. C=1
8. Else
9. C= 0.8
10. End if
11. End if
12. End for
13. End for

4. EXPERIMENTAL RESULTS

In order to evaluate proposed feature extraction results for web search engines results based on the domains like book, product, purchase products and auto. For each web database search engine, its LIS is build involuntarily using WISE i Extractor [15,14]. Some keywords are randomly selected from web search engine results to achieve the example result pages. The query terms are preferred in such a manner with the intention of they give way effect pages with numerous SRRs

beginning every one of WDBs of the corresponding domain. To measure the result of proposed PSO-DA (Particle swarm optimization with data alignment) using precision and recall measures beginning information retrieval. For data alignment algorithm with efficient feature extraction result from PSO, the precision is estimated based on the fraction of the appropriately associated data units greater than every one of the aligned units by the scheme; recall is the fraction of the data units with the intention of are appropriately aligned through the scheme over every one of physically aligned data units by the specialist. If the data unit is defined as appropriately annotated proposed system named label is exactly correct.

Precision is estimated based on the fraction of the appropriately associated data units greater than every one of the aligned units by the scheme it is defined as below:

$$Precision = \frac{t_P}{(t_p + f_p)}$$

Recall is the fraction of the data units with the intention of are appropriately aligned through the scheme over every one of physically aligned data units by the specialist it is defined as ,

$$Recall = \frac{t_P}{(t_p + f_n)}$$

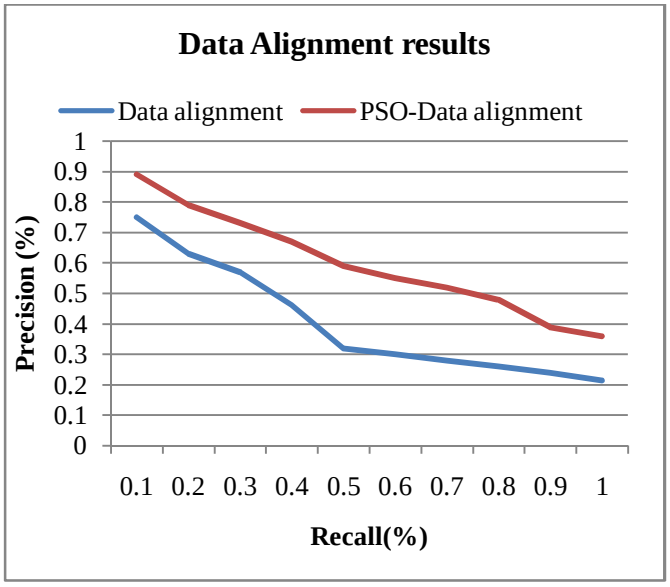

Figure 1: Data Alignment results

The Figure 1 shows the performance result between the data alignment and proposed PSO based feature extraction results from data alignment algorithm .The average precision and recall values of achieve higher results than earlier data alignment algorithm. Table 1 shows the performance result between the data alignment and proposed PSO based feature extraction results from data alignment algorithm .The average precision and recall values of achieve higher results than earlier data alignment algorithm. The performance of proposed PSO algorithm after feature extraction then alignment of data unit with semantically meaningful of each and every data units that achieved for each dataset.

Table 1: Data Alignment Results

Values	Data Alignment	PSO-Data Alignment
0.1	0.75	0.89
0.2	0.63	0.789
0.3	0.57	0.73
0.4	0.463	0.67
0.5	0.32	0.59
0.6	0.3	0.55
0.7	0.28	0.52
0.8	0.26	0.48
0.9	0.24	0.39
1	0.214	0.36

4. CONCLUSION

In this research proposed a PSO based feature extraction results for SRRs that extract individual text and data unit nodes separately. After important features are extracted from PSO algorithm for SRRs from web data base search engine .The feature names are labeled to each data unit and similarity between different data units are measured based on semantic similarity measure for each data units. Proposed system every self interested server (SIS) substitute data units are based on semantic similarity measure results using ontology and different ontology data unit's relationship procedures whereas difficult to achieve a compromise. The returned results as best data alignment and then perform annotation approach using wrapper methods for search results returned from whichever specified web database. Experimentation says that proposed PSO-DA optimization based feature extraction with efficient semantic similarity measurement best data alignment is helpful and they simultaneously are proficient of creating high-quality explanation of several web databases in the equivalent field.

References

1. Arasu and H. 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. Sixth Int'l Workshop the Web and Databases (WebDB),* 2003.
3. N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction,"

Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997.

4. 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
5. W. Meng, C. Yu, and K. Liu, "Building Efficient and Effective Metasearch Engines," *ACM Computing Surveys,* vol. 34, no. 1, pp. 48-89, 2002.
6. Z. Wu et al., "Towards Automatic Incorporation of Search Engines into a Large-Scale Meta search Engine," *Proc. IEEE/WIC Int'l Conf. Web Intelligence (WI '03),* 2003.
7. Cohen, W., Hurst, M., and Jensen, L." A flexible learning system for wrapping tables and lists in HTML documents",WWW-2002, 2002.
8. Kushmerick, N," Wrapper induction: efficiency and expressiveness", *Artificial Intelligence*, 118:15-68, 2000.
9. Lerman, K., Getoor L., Minton, S. and Knoblock, C. "Using the Structure of Web Sites for Automatic Segmentation of Tables." *SIGMOD-04*, 2004.
10. Pinto, D., McCallum, A., Wei, X. and Bruce, W. "Table Extraction Using Conditional Random Fields", *SIGIR-03.*
11. Buttler, D., Liu, L., Pu, C ,"A fully automated extraction system for the World Wide Web ", *IEEE ICDCS-21*, 2001.
12. Liu, B., Grossman, R. and Zhai, Y. "Mining data records from Web pages." KDD-03, 2003.
13. W. Bruce Croft ," Combining approaches for information retrieval", *In Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval,* Kluwer Academic, 2000.
14. V. Crescenzi, G. Mecca, and P. Merialdo ," Road RUNNER: Towards Automatic Data Extraction from Large Web Sites ", *VLDB Conference*, 2001.
15. S. Mukherjee, I. V. Ramakrishnan, A. Singh. "Bootstrapping Semantic Annotation for Content-Rich HTML Documents", *ICDE*, 2005.
16. J. Wang and F.H. Lochovsky," Data Extraction and Label Assignment for Web Databases", *WWW Conference*, 2003.
17. S. Greco, B. Matarazzo, and R. Slowinski, "Rough sets theory for multicriteria decision analysis," *Eur. J. Oper. Res.*, vol. 129, no. 1, pp. 1–47, 2001.
18. A. Skowron, "Rough sets in KDD-plenary talk," *in Proc. 16th World Comput. Cong. Intell. Inf. Process.*, 2002, pp. 1–14.
19. 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.
20. H. He, W. Meng, C. Yu, and Z. Wu, "Constructing Interface Schemas for Search Interfaces of Web Databases," *Proc. Web Information Systems Eng. (WISE) Conf.*, 2005.