Development of Web Annotation Technique for Search Result Records Using Web Database

Sonali T. Kadam[1],
Department of Computer Engineering,
S.B.Patil College of Engineering,Indapur,
University of Pune, India
Email ID:sonalitkadam@gmail.com

Sanchika Bajpai[2]
Department of Computer Engineering,
Bhivarabai Sawant Institute of Technology and Research,
University of Pune, India
Email ID: sanchi.scriet@gmail.com

Abstract **– This paper work investigates the needs which web users has to make annotations for their private as well as public use when they access the web pages.Web annotation has become an ongoing research issue since the invention of technologies such as HTML, XML, and Wiki.A Web annotation is an online annotation that deals with a web page.Semantic web and semantic data is related with integration scenario which defines the metadata construction methods.Web database helps to generate query result pages or search result records based on a user's query.A typical result page returned from a web database(WDB)has multiple search result records (SRRs). To get the web extracted records, SRRs to be a machine process able which is frequently useful for deep web data collection, online reading, and online shopping and have meaningful labels.In this paper, we have proposed exploratory automatic annotation techniques by making the use of automatically generated wrappers. Annotation technique firstly forms a group of data units on search result record such that same semantic will be assigned to data of a similar group by using multi-data alignment algorithm. Then we will annotate the every group which will be further aggregated to achieve the ultimate annotation label for the automatic wrapper generation purpose. These wrappers are used to generate annotations for new search result pages of the same database. Experiments will show that our proposed technique is efficient.**

Keywords: Web Data Extraction, Web database, Label Assignment, Wrapping, Annotation.

I. INTRODUCTION

Vast amount of information on the World Wide Web (WWW) is available since the search engines are becoming very much popular and important tools for people. According to the recent studies web searching is the second most favorite activity on the internet. Web users as well as web applications interacts with the search engines e.g. to perform the searching task meta-search engines [1] uses the available search engines and retrieves the dynamically generated search result records returned from available search engines. One more example is web crawling which is used to crawl data records and documents from deep web databases search engine [2].The representation and organization of the information items should provide the users

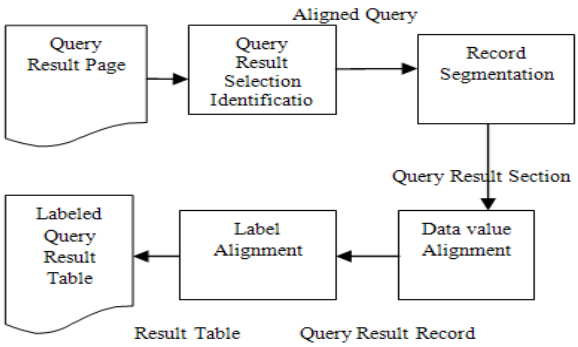

Fig. 1 flow diagram for data extraction

with easy access to information of their own interest. This paper gives a attention on major issues such as how to extract search result records(SRRs) or a query based search records(QBSR) from deep web databases, framework of metadata creation, assignment of labels to a text ,wrapper generation process, automatic annotation and data alignment algorithm.

Annotation plays a important role in our daily life and study. Annotation can be done for the Web, java, pdf, text, xps (XML Paper Specification), mobile, image, multimedia etc. Website which supports annotation systems, provide user friendly interfaces, annotation functions Website which supports annotation systems, provide user friendly interfaces, annotation functions. For many web site we suppose that these web sites will share the information.

We have noticed that while studying or reading text, reader needs more information or requires the text to be better explained and directly assigned to this word. So our goal is to study the annotation through which it will be possible to automatically add definitions or links to related pages to keywords in text of web pages. It will give the motivations to annotate the web database.

User query retrieves the relevant data which is in the form of HTML pages, structured or non-structured from web database. These HTML pages are called as the query result pages

978-1-4799-6892-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCUBEA.2015.73

.Automatic information extraction is vital for meta querying, data integration, big data and data warehousing only when data extraction is done using querying technique. Data extraction flow [3] is shown in Fig.1.

1. *Query result section identification: It* describes how much portion in a dynamically generated query result page (SRRs) contains the data that need to be retrieved.
2. *Record segmentation : S*egments the query result section into different records and extracts out them.
3. *Data value alignment:* Data values are aligned from various records that comes under the same attribute so that they can be grouped and arranged into a table form.
4. *Label assignment: Label* assignment assigns a suitable, meaningful label to each column in an aligned table.

In this paper our proposed work is to annotate the database by studying the multi data alignment problem. Clustering based shifting machine learning technique and more sample pages from each training site will be obtain to identify the efficient automatic annotation technique for the web annotation. This paper has proposed the following contributions.

1. Mostly existing approaches simply design labels to HTML nodes, here we want perform the data level annotation.
2. Multi data alignment algorithm will be used to group the data units that have same semantics by using clustering based shifting technique.
3. Integrated interface schema (IIS) can be used for SRRs annotation.

Our proposed multi-data alignment method ([4], [16])consists of three stages as shown in Fig.2.Fisrt stage is a multi-data alignment stage in Fig.2a) identifies the data units in SRRs and then arranges these into various groups as shown in Fig.2b). Each group will have same semantics which helps to find out common patterns and features within these data units. Secondly in annotation stage query based annotation is used to assign meaningful labels for the records within a group as shown in Fig. 2c). In third annotation wrapper generation stage as shown in Fig.2d) for each concept we will create annotation rule that describes how the data records are extracted in a search result records and assignment of exact semantic label. These rules are for all multi data aligned groups which forms the annotation wrapper for WDB. Then these wrappers are directly used to annotate the data retrieved from same WDB in reply of new queries.

This paper provides the contribution in literature survey which introduces some of the currently existed and similar annotation systems with their pros and cons, implementation details such as design of our paper, system architecture, data alignment algorithm, and platform. Further paper describes the conclusion and future scope of this area.

II. LITERATURE SURVEY

The literature survey includes several sections each representing important inventions done by the researchers for the web annotation. Important areas to be discussed are the annotations, related technologies, existing systems, their pros and cons. Each section is demonstrated as follows.

A. *Bidirectional Interaction between the Users*

The recent structure of WWW has the bounded ability of users to do the communication with web environment as well as with individuals on Web [5]. Now many technologies and systems have been investigated for the interaction and collaboration between users such as wiki, blogs, social networking websites etc.

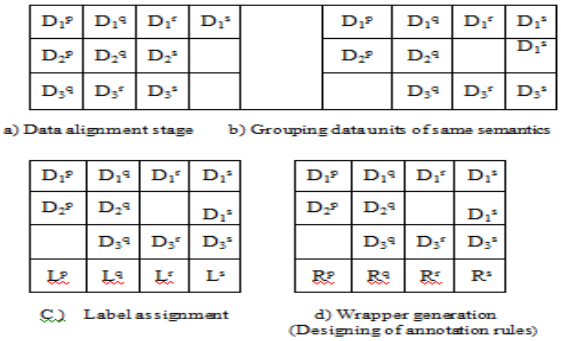

Fig. 2. Explanation of our three-stage annotation solution

B. *Web Annotation*

Web annotation and annotation process is described as follows.

1) Annotation: Annotation is an interesting research issue since the invention technologies such as HTML, XML, Wiki, JAVA, IT based, content-based image retrieval (CBIR) etc. In this paper, we have defined the web annotations [6] as: "Online annotations associated with web resources such as web pages, with which users can add, update or delete a text, image, comment which include highlights or underlining, footnotes, tags, and links from a web page without modifying the page itself". Annotation can be done for the Web, java, pdf, text, xps (XML Paper Specification), mobile, image, multimedia etc[8].Annotation can be classified into different types such as web annotation, IT based annotation, JAVA annotation, text annotation, mobile annotation([6], [9], [10]).To annotate the data different annotators are used such as Table annotator, Query-based annotator, Schema value annotator, Frequency-based annotator, Same-prefix annotator, Common knowledge based annotator as explained in [11].Annotator is designed for Netscape communicator.

2) *Annotation Process*: An annotation process [13] of server-side markups or tag is supported by the user interface as follows:
Input: A Web site retrieved from web database.

1. The browser helps the user to open a server-side marked up web page which is produced by database owner.
2. The browser handles the server-side markup , e.g. it provides graphical icons on the web page, so that identification of values which come from a database becomes easy.
3. The annotator produces client-side annotations after doing the confirmation with client and server side markup.
4. Mapping rules and annotator's ontology are available on the web. The user chooses the server-side markups to map the rules.
5. The querying party side retrieves the results from database.

The annotator can annotate an organization entry from web pages. Then, annotator can use the ontology and mapping rules to instant services by regularly querying for all current title entries from the collection of topics.

C. *Existing Annotation Systems and Comparison*

Data extractors utilize regular expression which are explained in [3].There are the different techniques on data extraction [4]which were introduced by researchers. A novel data extraction technique ODE(Ontology Assisted Data Extraction) which retrieves the query result record (QRRs) or SRRs automatically from the web database(WDB). The different techniques such as ROADRUNNER ,WISEiExtractor, ViDE, Novel Schema model, DeLa, Partial Tree Alignment are briefly explained by various researchers in [4]. Annotation systems are consists of three modules [5] such as to view existing annotation, to create new annotation and to save the annotations. Four type of relationships are identified in [11] that exist in the information and that are one to one, one to many, many to one and one to nothing relationships. CREAM [15] allows the creation of metadata which includes authoring mode. In authoring mode, authors create the metadata or relational metadata. ViDE [16] is used to extract the data from deep web database. Wrapper induction systems sometimes are not suitable for web application ([21], [22]) since it has to extract information from large number of web pages.

III. IMPLEMENTATION DETAILS

A. *Mathematical Model*

Text node features are like data content, Presentation Style, Data Type, Tag Path, and Adjacency etc, data alignment includes data unity similarity which include Data content similarity (SimC), Presentation style similarity (SimP), Data type similarity (SimD), Tag path similarity (SimT), Adjacency similarity (SimA).

Data Units: Each node in a tag structure is either a tag node or a text node. A tag node corresponds to an HTML tag surrounded by "<" and ">" in HTML source, while a text node is the text outside the "<" and ">." Text nodes are the visible elements on the webpage and data units are located in the text nodes.

1. *Data Units and Text Node:* We identify and use five common features shared by the data units belonging to the same concept across all SRRs.

1.1 Data Contents: The data units or text nodes with the same concept often share certain keywords. This is true for text nodes that contain data units of the same concept usually have the same leading label.

1.2 Presentation Style: It describes how a data unit is displayed on a webpage. It consists of six style features: font face, font size, font color, font weight, text decoration like underline, strike, etc., and whether it is italic.

1.3 Data Type: Each data unit has its own semantic type although it is just a text string in the HTML code. The basic data types are date, time, currency, integer, decimal, percentage, symbol, and string which currently considered in our approach.

1.4 Tag Path: A tag path of a text node is a sequence of tags traversing from the root of the SRR to the corresponding node in the tag tree. Each node in the expression contains two parts, one is the tag name, and the other is the direction indicating whether the next node is the next sibling (denoted as "S") or the first child (denoted as "C").

1.5 Adjacency: For a given data unit d in an SRR, let d^p and d^s denote the data units immediately before and after d in the SRR, respectively. We refer d^p and d^s as the preceding and succeeding data units of d, respectively. Consider two data units d_1 and d_2 from two separate SRRs. It can be observed that if d_1^p and d_2^p belong to the same concept or $d_1^s$and d_2^s belong to the same concept, then it is more likely that d_1 and d_2 also belong to the same concept.

2. *Data Alignment*

2.1 *Data Unit Similarity:* The purpose of data alignment is to put the data units of the same concept into one group so that they can be annotated holistically. In this paper, we have represented the similarity between two data units (or two text nodes) d_1 and d_2 is a weighted sum of the similarities of the five features between them, i.e.:

$$Sim(d_1,d_2)=\omega_1*SimC(d_1,d_2)+\omega_2*SimP(d_1,d_2)+\omega_3*SimD(d_1,d_2)+\omega_4 SimT(d_1,d_2)+\omega_5*SimA(d_1,d_2) \quad (1)$$

The similarity for each individual feature is defined as follows:

Data content similarity (SimC). It is the Cosine similarity between the term frequency vectors of d_1 and d_2:

$$SimC(d_1,d_2)=(V_{d1}*V_{d2})/(||V_{d1}||*||V_{d2}||). \quad (2)$$

where V_d is the frequency vector of the terms inside data unit d, $||V_d||$ is the length of V_d, and the numerator is the inner product of two vectors.

Presentation style similarity (SimP). It is the average of the style feature scores (FS) over all six presentation style features (F) between d_1 and d_2:

$$SimP(d1,d2) = \sum_{i=6}^{6} FSi / 6 \quad (3)$$

Where FS_i is the score of the *i*th style feature and it is defined by FSi = 1 if $Fd_1^i = Fd_2^i$ and $FS_i = 0$ otherwise, and Fd_i is the *i*th style feature of data unit *d.*

Data type similarity (SimD). It is determined by the common sequence of the component data types between two data units. The longest common sequence (LCS) cannot be longer than the number of component data types in these two data units. Thus, let t_1 and t_2 be the sequences of the data types of d_1 and d_2, respectively, and TLen(*t*) represent the number of component types of data type *t*, the data type similarity between data units d_1 and d_2 is

$$SimD(d_1,d_2)=(LCS(t_1,t_2))/(Max(Tlen(t_1),Tlen(t_2)). \quad (4)$$

Tag path similarity (SimT). This is the edit distance (EDT) between the tag paths of two data units. Let p_1 and p_2 be the tag

paths of d_1 and d_2, respectively, and PLen(p) denote the number of tags in tag path p, the tag path similarity between d_1 and d_2 is

$SimT(d_1,d_2)=1-(EDT(p_1,p_2))/(PLen(p_1)+PLen(p_2))$. (5)

Adjacency similarity (SimA). The adjacency similarity between two data units d_1 and d_2 is the average of the similarity between d_1^s and d_2^s and the similarity between d_1^p and d_2^p is

$SimA\ (d_1,d_2)=(\ Sim'(d_1^s,d_2^s)+\omega_2*Sim'(d_1^p,d_2^p)\ )\ /\ 2.$ (6)

When computing the similarities (Sim') between the preceding/succeeding units, only the first four features are used. The weight for adjacency feature (ω_5) is proportionally distributed to other four weights.

3. *Annotation Wrapper:* Data records are annotated on a result page , we will use these annotated data units to build a wrapper. Our automatic wrapper generation method [17]is as described here. Wrapper is the annotation rules for all attributes on retrieved result page. Annotation rule for attributes consists of parameters such as Attribute = < label, prefix, suffix, and index >. Scan the group in bidirectional way to get prefix and suffix of data unit. These wrappers are used to annotate new result page of same WDB.

4. *Assigning Labels:* Retrieved data has the attributes $=(A_1,A_2,..A_k)$.Return SRRs is always related with query and query item can be entered into textbox or can be selected from selection list on a interface.. By using proper, structured, context information mapping rules can be designed i.e. ontology [3].

B. Design

In web annotation paper we have designed modules such as a client side module which helps for authorized registration and operations to be performed, web database to get the search result records i.e. data extraction after sending the queries by users, wrapper generation module which includes the multi-data alignment by using clustering based technique for the wrapper generation for a data, annotation server which annotates the data on web pages of search result records and finally these web pages will be displayed to see the web annotation results.

C. Architecture

The basic architecture of web annotation ([13], [15]) is depicted in Fig. 3. The Design pursues the idea to be flexible and open for the implementation of the web annotation. A plug-in is used to insert or replace modules in paper.GUI of web annotation consists of document viewer/editor guidance and fact browser together. Plug-in that can establish new connections to add new modules in a paper and one can replace a module also. The design consists of GUI for annotation, document management, annotation server, information extraction from web database and automatic wrapper generation etc.

D. Multi-Data Alignment Algorithm

Our multi-data alignment method ([4], [16]) assumes that attributes appear in a same order into the SRRs but SRRs actually contains various attributes. Alignment group is an each table column which consists of at least two data records. Goal of our algorithm is to do the well data alignment in which same column or aligned group which are of same semantics data items are placed.

Our multi-data alignment algorithm is shown in Fig. 4 and steps are mentioned as follows:

1) Combine the matched items: This step detects combines the matched items into a single node that has a same attribute.

2) Align the matched items: This step aligns the matched items into group and whose concepts are same.

3) Assign null value to blank item: In this step empty data items will be fill up by null value.

4) Align data records: Each group is converted into the record and each record is having data of same concept.

The input is n data records $\{R^1,R^2,\ldots R^n\}$which are extracted from WDB. Data record Ri, is a sequence of data items $(D_i^1,D_i^2,..D_i^k)$.Data item has unique position in its sequence. Time complexity of this algorithm is $O(n^2*k)$ where n is total data records and k is average number of total data items per record.

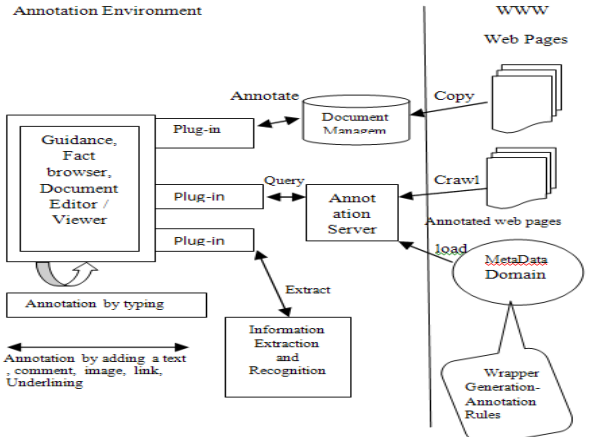

Fig. 3 Architecture of web annotation

Multi-data alignment algorithm:

Input –Group of extracted data records $\{R_j|1<=j<=m\}$

Output-Group of data records $\{R_i|1<=j<=m\}$ including all aligned data.

Start

1.RecentItemSet=NULL;

2.RecentCluster=NULL;

3. RecentItemSet $=D_j^{U(j)}(1<=j<=m)$;

4.while RecentItemSet=NULL;

5.MatchDataItem();

//to group RecentItemSet in x ,cluster $(Cj|1<=j<=x);(x<=m)$

6.for each cluster Cj

7.for each Ry

8.if $D_y^{U(y+x)}$ is matched to Cj

9.log position x;

10.else log position 0;

11.A_j = Max(log position) for Cj; //A_j is position of item

12.RecentCluster==C_L

13.else RecentCluster= C_L whose A_L =Max$\{A_1,A_2,..A_x\}$

14.for R_j delete $D_j^{U(j)}$ from RecentItemSet

15.if $D_j^{U(j)+1}$ present in R_j then put it in RecentItemSet

16.for each R_j move NULL item in front of $D_j^{U(j)}$
17.U(j)++;
End
MatchDataItem()
Input-i1, i2 // data items Output- Matched result
Start
1. if(front(i1)!=front(i2)) Then return unmatched;
2. if(pos(i1)=pos(i2)) then Return matched;
3. if(i1 and i2 matched) then Return matched;
Else return unmatched;
End

Fig 4.Multi-data alignment algorithm

IV. RESULTS

Collected WDBs are randomly divided into different groups. Data set is formed by obtaining one sample result page from each training site. Data sets are generated by collecting two sample result pages from each testing site using different queries. For each result page in this data set, the data units are manually extracted, aligned in groups, and assigned labels by us. Here we can add a note, highlight and modify the data, also we can do the image annotation. In Figure 4 notes are added and highlighted the data on result page from web database.

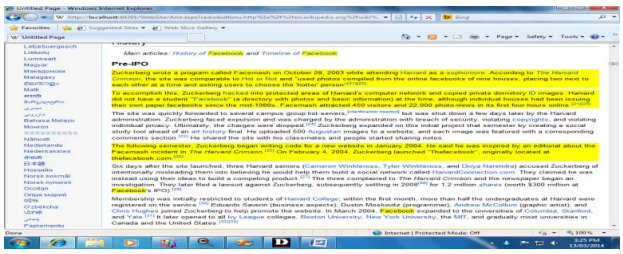

Fig. 4 Annotation Results.

V. CONCLUSION

With the flourish of web database users can have the golden opportunity to benefit the web annotation. In this paper we have focused the data annotation and multi-data alignment problem. We proposed an annotation approach to enhance the annotation for search result of query result records retrieved from web database. This approach includes multi-data alignment method and query based annotation. Our experimental results will show that how annotation is useful for users and how this is an efficient technique to enhance the annotation methods. To get accurate annotation we will use the clustering based shifting method in this paper. Here we have used more sample pages from WDB for the annotation.

However, there are more improvements areas are having the chances in different annotation types such as web annotation, java annotation, IT based annotation [6], PDF annotation, Multimedia (audio, video, image) annotation ([18], [19], [20]), XPS annotation, text annotation. Instead of using only some samples pages from web database (WDB) the deep web database can be used by using machine learning techniques.

For coming future work, there is a huge list of open issues concerning web annotation like granularity, Automatic derivation of server-side web page markup, Other information structures, Inter-linkage, query based database can create the security issues such as denial of service attacks as discussed in the conclusion of [13].

REFERENCES

[1] W. Meng, C. Yu, K. Liu. "Building Efficient and Effective Meta- search Engines", *ACM Computing Surveys*, 34(1),March 2002, pp.48-84.

[2] S. Raghavan, H. Garcia-Molina. Crawling the Hidden Web.*VLDB Conference*, Italy, 2001.

[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] Yiyao Lu, Hai He, Hongkun Zhao, WeiyiMeng, and Clement Yu, "Annotating Search Results from Web Databases" *IEEE Trans. Knowledge and Data Eng., vol. 25, no. 3, pp. 514-527, Mar. 2013.*

[5] PeiwenZhu,"Web Annotation Systems: A Literature Review and Case Study ",A Master's Paper for the M.S. in I.S. degree ,Chapel Hill, North Carolina pages 35,April, 2008.

[6] http://en.wikipedia.org/wiki/Annotation

[7] Xin Fu, Tom Ciszek, Gary Marchionini, Paul Solomon, "Annotating the Web: An Exploratory Study of Web Users' Needs for PersonalAnnotation Tools" ,*in The 68th Annual Meeting of the American Society for Information Science & Technology (ASIS&T)*,Charlotte,NC,USA,2005

[8] Ng S. T. Chong," Annotation-based Web Communications Systems: A Review",*Technical report CS-3408*,Oct-2003

[9] Vincenzo Gervasi, Giacomo A. Galilei, "Software Manipulation with Annotations in Java", *Advances in Software Engineering*,LNCS 5316, , Springer,pp 161-148,2008.

[10] S Feng, R Manmatha, and V Lavrenko"Multiple Bernoulli relevance models for image and video annotation". *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1002–1009,2004.

[11] Yiyao Lu, Hai He, Hongkun Zhao, WeiyiMeng, and Clement Yu, "Annotating Search Results from Web Databases" *IEEE Trans. Knowledge and Data Eng., vol. 25, no. 3, pp. 514-527, Mar. 2013.*

[12] Zohar, R." Web Annotation - an Overview",*Dept. of Electrical Engineering, Israel Institute of Technology*,1999.

[13] S. Handschuh, S. Staab, and R. Volz, "On Deep Annotation",*Proc. 12th Int'l Conf. World Wide Web (WWW)*, 2003

[14] http://en.wikipedia.org/wiki/Web_annotation

[15] S. Handschuh and S. Staab, "Authoring and Annotation of Web Pages in CREAM," *Proc. 11th Int'l Conf. World Wide Web (WWW), 2003.*

[16] 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.*

[17] [17] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction", *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997.*

[18] S Feng, R Manmatha, and V Lavrenko"Multiple Bernoulli relevance models for image and video annotation", *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1002–1009,2004.

[19] G Carneiro, A B Chan, P Moreno, and N Vasconcelos ,"Supervised Learning of Semantic Classes for Image Annotation and Retrieval". *IEEE Trans. on Pattern Analysis and Machine Intelligence*. pp. 394–410,2006.

[20] Fusheng Wang ,Rabsch, C. , Peiya Liu, "Native Web Browser Enabled SVG-based Collaborative Multimedia Annotation for Medical Images" ,*Data Engineering*, Pages 1219 - 1228 ,ICDE 2008

[21] Z. Wu et al., "Towards Automatic Incorporation of Search Engines into a

[22] Large-Scale Metasearch Engine," *Proc. IEEE/WIC Int'l Conf. Web Intelligence (WI '03), 2003.*

[23] W. Meng, C. Yu, and K. Liu, "Building Efficie nt and Effective Metasearch Engines," *ACM Computing Surveys, vol. 34, no. 1,pp. 48-89, 2002.*