RESEARCH ARTICLE OPEN ACCESS

Review of Semantic Web, Annotation Methods and Automatic Annotation for Web Search Results

Shilpa Jadhao[1], Prof. R. P. Kulkarni [2]

[1](Department of Co mputer Engineering, Pune University, India Email: shilpa.jadhao@mctrg it.ac.in)
[2](Department of Informat ion Technology, Pune University, India Email: rpk.sit@sinhgad.edu)

ABSTRACT
Now days the use of web search engines is very frequent and common worldwide over the Intern et by end users for different purposes. The basic aim of this web search engines is that to take the query request from the end user and execute that query on relational database used to store the information on behalf of that web search engine. Based on input queries the dynamic response is generated by search engine. There are many databases available those are supporting the HTM L through web accessibility for their data. Whenever end user submits their query for searching, the dynamic web pages extracted as result for that query. Every web page which is generated is containing the many results to display for particular query. The result of query is called as Search Result Records (SRRs). These SRRs containing many data points those are relevant to one common semantic. SRRs further required to be assigned with proper labels. The manual methods in which records are extracted fro m search results and then manually labels are assigned to them, however this method is having worst scalability. Thus the new automatic annotation based methods presented recently to improve the accuracy as well as scalability of web search engines. In this paper we are taking the review of such systems. In addition to this we are discussing the semantic web using rational databases.

Keywords – Annotation, Relat ional database, scalability, Semantic web, SRRs, web search engines.

1. INTRODUCTION

Since fro m the last two decade, the concept of semantic web is interestingly becomes the area of research for many researchers. The main inspiration of this concepts introduced by scientist Tim Berners Lee. For designing of Semantic Web, many tools, languages, and methods introduced by many authors and still mo re research is going on to make it more robust, faster and efficient over the web. The semantic web is basically based on use of relational databases in order to serve the services to end users based on their requests [1]. Fro m the in itial phase of semantic web, the use of relational databases and their ro le was investigated due to reason that relational database at first compared with global database as well as it is relat ional database was relative new concept in database field. But the exchange and collaboration of the ideas among these did not unidirectional. As the research progresses, there were many innovations introduced by various researchers in order to make semantic web based search engine efficient [8]. W ith the introduction of semantic web, many business portals looking Web as best way of informat ion presentation. Web search engines are most commonly used now days and the result of queries over web search engine generated as dynamic web pages. The data which is resulted is initially unstructured and unlabeled, thus to do the automatic labeling to data points automatic annotation methods was introduced. This annotation and alignment of SRRs data improves the searching efficiency as well as updating data.

The process of data alignment is nothing but data arrangement and its accessibility using the computer memo ry. The process of data annotation is method of inserting the data into the web document semantically. This process provides the immed iate extraction of data from the deep web. As we discussed in abstract also the results retrieved from database is called as search result records (SRRs)

based on input user queries. Every SRRs are consisting of different data units.

The data units from the SRRs are dynamically encoded into the search retrieved web pages for the sake of end user browsing as well as translate into the mach ine reading unit with the assignment of the mean ingful labels. The manual process of labeling to the extracted data units requires more time as well as less scalability and hence less accuracy of search results. Thus to overcome the limitations of existing methods, the recent automatic annotation methods was introduced by various authors. This automatic method improves the scalability as well as accuracy of search engine. In this paper we are presenting review of semantic web and role of relational databases, review of document annotation methods, and finally the review of automat ic annotations for web databases.

2. REVIEW OF ANNOTATION METHODS

The annotation definition varies according to the application domain is worth noting: [2] Linguistics, biology, e-learning [3] [9] [7] and Web application software development [10]. Ho wever, it can be said that all graphic or text annotation is attached to an informat ion document, including documents, various institutions; it is a document, a path, a sentence, a Word, a word or an image [7] refers to a set of many methods and techniques to exp lore the relevant annotation systems[8], conceptual graphs [3], meta-thesauri [9] and linguistic ind icators [4] are used to present as we describe, Thereafter the main existing annotation systems. SyDoM [3] is a semantic annotation of Web pages system. Th is allo ws the enrich ment of these pages so that they take account of their writ ing without language find it with textual XM L format is dedicated to manage documents stored. we see that SyDoM has two main advantages: first, mult ilingual research and other But the improvement of the representation of Web pages, we SyDoM out research on Web pages only if it has been already annotated, yet this annotation by us ing different means to inquire Web pages created thesauri have been unable to get that information. EXCOM [12] is an annotation engine internal/external annotating a document by a clique of hay knowledge on aim uses a set of linguistic devices. This engine is under development, and at the present time, a cosmic stories and questions allows the production of an automated reformu lation. However, we observe that this is still an important part of the system implemented, i.e. by considering meaning documents annotated informat ion indexing.

Annotea [12] a co llaborative client server system document annotation is a special they are stored on the server in such a way that anyone who has access to an annotation server for a given document to consult all related annotation and add your own annotations will be enabled for these annotations are divided into typing comments Improve projections, assumptions. This system was developed using W3C standards. Yet, only possible Co mmittee on State annotation text; It is annotated by a picture or symbol.[9] Acacia team allows annotation system developed by genes. This creature, which experiments to validate and to interpret the results, obtained on the biopuces helps make system research difficult task them. Its genetic database offers the possibility of a key word research. For key word can or a biological phenomenon Jean correspondence study. All previous works are interested in general documents annotation like scientific articles, Web documents, biological databases and mult imedia documents. Only few of them focus on the events annotation. We present, in the following, some of these works: The annotation of temporal informat ion in texts [9]: this work focused more specifically on relations between events introduced by verbs in finite clauses. It proposes a procedure that achieves the task of annotation and a way of measuring the results. The authors of this work tested the feasibility of this procedure on newswire articles with promising results. Then, they developed two evaluation measures of the annotation: fineness and consistency.

Annotating texts [5] features and relat ionships to determine the relative annotation scheme: This enables order and, if possible, absolute time events. A planning an annotated corpus can be used for building the corpus is usually producing such benefits associated with building resources. It also can be used to better understand the phenomena. Plus it training and adaptive algorithms for evaluating represents a source. It automatically shows the relationship of the features and interest. However, we observed that the relationship between the incidences of this work to determine based on temporal markers only. There are inherent differences with regard to events without using temporal markers which are

expressed, since this technology is not entirely accurate.

3. REVIEW OF ROLE OF DATABASES IN SEMANTIC WEB

A Semantic Web approach has many benefits and the importance of the database to ontology mapping can be used in a database use cases clear. After all, the problem of semantic web mapping in a database other than a representation model transition as a mere exercise of It did not emerge in d ifferent motivations and goals and challenges to be successful a clear separation of SW technologies, relational databases and interactions between imp licat ing is important to identify problems. There are quite a few versatile tools that two (or mo re) birds with one stone to kill corruption as methods and approaches presented here responds to a special case of strictly That is not to say that using later, we will have some benefits that databases and ontology's can be obtained from the inter connection present. Dynamic Web pages semantic annotation aspect the Semantic Web vision of changes to the current Web a Web documents. To achieve this in a direct way to annotate HTML pages, which presented the way your content and are only suitable for human consumption for specifying ontology's semantically HTM L pages. Software by agents and their content, Web services for processing suite enabling words can be annotated with RDFa (Resource Description Framework in Attributes) in XHTM L that embeds reco mmendation. Proposed terms of reference ontology tag considerably since such annotation facility. However, these dynamic Web pages that retrieve their content directly to the underlying database quite well doesn't work: this one, wikis for content management system (CMS), case and other Web 2.0 sites. Represent World Wide Web dynamic Web pages, the so-called deep Web, since the creation of these pages to a Web service or Web form interface is generated in response to the request which search engines and software agents, is not accessible to largest. Insofar As the Web page owner is willing to reveal the structure of your database it every dynamic page manual annotation of infeasibility, one possible solution "for" annotate directly to the underlying database schema, has argued that the "," database schema annotation elements and dynamic page content fits a previously existing domain ontology is a set of correspondences between. Once such mappings are defined it's embedded in automated fashion content derived annotations with Generate dynamic enough to pinpoint the semantically annotated pages.

Heterogeneous database integration resolution of database research area, the diversity that the remains, to a large degree, the most popular, long unsolved issues. Diversity occurs when different software or hardware infrastructure, follo w print different syntactic conventions, use two or more database systems or between when they interpret differently the same or similar data. The resolution of the above forms of diversity to be grated mult iple databases and their contents to be uniformly allo wed queried. Typical database integration architectures, conceptual models in one or more description of the contents of each source database used to live against a global conceptual schema. Queries are generated; one for each source database to retrieve the appropriate data query and reformulate the wrapper is responsible for integrating ontology's ontology -based conceptual schemas. And, therefore, to be among the source databases are emp loyed in lieu of correspondences and define to be one or more ontology's. Conditions of such a source database of correspondences ontology-view or LA V mapping (local), exp ressed as a conjunctive query against the source mapping express ontology conditions source database (as a global view or GA V mapping) as a conjunctive query against either or both the source database and ontology (global local scene or GLA V mapping) is an equivalence of two queries against the Kingdom. Mapping used in an integration architecture affects the type of query processing complexity and extensibility, both of the whole system (e.g. in the case of query processing GA V mappings is insignificant, but a new addition to the source database requires a reinterpretation of all mapping, the inverse holds for LA V mappings). Thus, discovery and relational database schemas and ontology have to represent the mapping between heterogeneous database integration scenarios constitute an integral part. Much like a database reaches the ontology based data integration architecture, ontology-based data access (OBDA) assumes that an ontology is a source database, and the data collected between acting as an intermed iate layer is lin ked to a OBDA system objectively end which underlying data source obscure storage details do not need to know about the user a notification

system to offer high-level services . In terms of interest Ontology database, queries a domain to a high level of detail. A llo w the user to an intangible area. in the up-per hides the details data source-specific levels of local data source schema queries against a conceptual changing queries against in some way, an information integration scenario in which a cover resembles the OBDA engine query rewrite OBDA en-account mapping in a database and a contextual ontology to describe the domain interest in the midst of taking gene is performed by a main benefit of OBDA architecture is the fact that meaning in RDF queries without the need to replicate the entire con tent a are generated directly against the database.

In addit ion, OBDA applications ontology mapping where a database output to a reformulated SQL query user intended to better capture the semantic SQL queries can be useful to rewrite the rewrit ing related by substituting synonyms and ontology with the terms and conditions of use of the original SQL query is performed by the application associated query relational data reference another remarkable external ontology's to use as It has the ability to feature some database management sys - SQL queries in terms o f prior conditions including a pressed-allowing ontology terms, have been implemented. Mass generation of Semantic Web data, it has been argued that due to the delay in receipt of Semen TIC Web managed tools and applications performance SW technologies advantages. Such equipment, however, the success of SW data, a "chicken and egg problem" leading to the availability of a sufficiently large volu me correlated directly to where cause and effect circle form a vicious. relat ional database is one of the most popular storage media on the World Wide Web are holding most of the data, since generation SW data would be a solution to a critical mass, preferably automatic withdrawals relat ional database content RD In the SWF data, soft ware developers and equipment manufacturers will create a significant pool of inhibit ions, and in turn, the increased production anticipated to be SW applications. The ontology mapping database for literature the term transformations as well has been used to describe. Manually learn the process of developing ontology fro m scratch ontology is hard, time consuming and error p rone. Many semi automat ic ontology teaching methods, free and semi structured text documents, vocabularies and thesauri, domain experts and other sources knowledge extract ing mot ion. Relational database structured information sources and, in case their schema (i.e. a conceptual models, such as UML or extended entity relationship model based on design) modeling the following standard practices, they domain knowledge is important. The format ion of national sources where enterprise databases are maintained and often time data especially in business environments is true. so, rich ontology's relational database until a do main expert monitor process and the end result of learning enriches the informat ion their schemas, content, queries and stored procedures, may be removed by the House from the learning a common ontology inspiration. When the particular domain interest in a situation is existing ontology database ontology is driving for the mapping that frequently is not so many years ago. However, as the years passed by, ontology learning technique is primarily a wrapping ontology source relational database to access data for ontology based or database integration is used to create a reference.

Defin ition of the intended meaning of a relational schema as already mentioned, standard database design practices begin with the design of a conceptual model, which is then transformed, in a step known as logical design, to the desired relational model. However, the in itial conceptual model is often not kept alongside the implemented relational database schema and subsequent changes to the latter are not propagated back to the former, while most of the times these changes are not even documented at all. Usually, this results in databases that have lost the original intention of their designer and are very hard to be extended or reengineered to another logical model (e.g. an object-oriented one). Establishing correspondences between a relational database and an ontology grounds the original meaning of the former in terms of an exp ressive conceptual model, wh ich is crucial not only for database maintenance but also for the integration with other data sources, and for the discovery of mappings between two or mo re database schemas. In the latter case, the mappings between the database and the ontology are used as an intermed iate step and a reference point fo r the construction of inter database schema mappings.

4. REVIEW OF METHODS OF AUTOMATIC ANNOTATIONS

Consider a set of SRRs that are extracted from a

result page returned from the web database. The Automatic annotation approach has three major phases as illustrated in Figure. 1. Let d_i^j denote a data unit, belonging to the i^{th} SRR of concept j. Figure 1a represents SRR in table format [12].

- Align ment phase [6]: First data alignment phase in the SRRs units identified and organized into different groups for each group corresponds to a different concept (for examp le, all titles of books are grouped together). Figure 1b across all SRRs step 1 each column containing data unit with sane sense results. This step is to identify the patterns and features of data between units are used.
- Annotation phase [6]: Annotation phase each with many basic annotators features one type of exploitation. Every annotator groups organized within data units of a label that is used to determine the most suitable probability models a used to label. Figure 1 shows the results of step 2 c where a meaning with each group assigned labels l.
- Annotation wrapper generation [6]: Annotation wrapper generation phase one annotation rules for each identified entity or concept RJ has generated the data unit description, how to remove and what means should be labeled and collectively cover a cover forms a new queries for data retrieved fro m the Web database units are used to annotate and thus annotations quickly.

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b	d_2^d	
d_3^b	d_3^c	d_3^d	

(a)

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b		d_2^d
	d_3^b	d_3^c	d_3^d

(b)

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b	d_2^d	
d_3^b	d_3^c	d_3^d	
L^a	L^b	L^c	L^d

(c)

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b	d_2^d	
d_3^b	d_3^c	d_3^d	
R^a	R^b	R^c	R^d

(d)

Fig 1 Illustration of three phase annotation approach [6]

4.1. Data unit and text node [6]*:* The visible elements on the web page represent a text node and the data units are located in the text nodes. Relationships between text node and data unit features are,

- One-to-One Relat ionship**:** Text node containing exactly one data unit, i.e. the text of this node contains the value of a s ingle attribute. Each text node surrounded by the pair of tags <A> and </A>. This type of text nodes are referred as atomic text nodes. An atomic text node is equivalent to a data unit.
- One-to-Many Relationship: Multiple data units are encoded in one text node. This type of text nodes are referred as composite text node.
- Many-to-One Relat ionship: Multiple text nodes together form a data unit. Th is type of text nodes is referred as decorative tags because they are used for changing the appearance of part of the text node.
- One-To-Nothing Relationship: Text nodes are not part of any data unit inside SRRs. This type of text node is referred as template text node. There are five co mmon features shared

by the data units Data content Presentation style Data type Tag path Adjacency

4.1.1 Data content [6]: Data unit or text node of same concept shares certain keywords wh ich are used to search the informat ion quickly. For e.g., keyword "mach ine" will return the informat ion that are relevant to word machines.

4.1.2 Presentation style*:* Presentation feature describes how a data unit is displayed on a web page. Few of the styles are font face, font size, colour, text decoration etc.

4.1.3 Data type: Data types are predefined characteristics that have their own meaning. Basically used data types are date, time, currency, integer, decimal etc.

4.1.4 Tag path: A Tag path is a sequence of tags traversing fro m the root of the SRR to the corresponding node in the tree. Each node contains two parts a tag name and a direction indicating whether the next node is a sibling or the first child node.

4.1.5 Adjacency*:* Adjacency refers to the data units that are immediately before and after in the SRR. They are termed as preceding and succeeding data unit.

4.2 Data alignment and labeling [6]: Current tasks when compared with automatic annotation approach. They are based on one or a few facilities. Automatic annotation alignment approach first data units and handles relations between text nodes and data unit do use variety of features the device is a cluster-based transfer algorith m and is used in the alignment process. Label assignment IIS (unified interface schema) and LIS (local interface schema). There are attributes in all LIS IIS and thus eliminates inadequacy and inconsistent labels label problems. In the coalition of so me basic annotators groups started to annotate and combine mult iple annotators a probability model is used for the results of this approach are called mu ltip le-annotator approach.

5. DATA EXTRACTION METHOD COMPARISION

Method	Nested Structure Processing	Single Result Page	Non-contiguous Data Regions
CTVS	✓	✓	✓
DeLa	✓	✓	×
ViPER	×	✓	×
ViNTs	×	×	×

Table 1. Data extraction methods comparison

6. CONCLUSION AND FUTURE WORK

In this survey paper, we have introduced the concepts of semantic web in details. The ro le and motivation of using the relational databases into the semantic web discussed. The semantic web concepts are base for the web search engines, therefore our survey first conducted over the same. After that we introduced the different annotation techniques present for documents. Later these annotation methods referred for the automatic annotations of web search results in web search engines for improving the scalability and accuracy. This paper was prepared by considering our next research work over the same research domain and problem. For the future we suggest to propose efficient annotation of web search databases with its practical analysis

ACKNOWLEDGMENT

Thanks to the experts who have contributed towards the development of this material.

REFERENCES

Journal Papers:

[1] Ay men Elkhlifi and Rim Faiz, "Automatic Annotation Approach of Events in News Articles", *International Journal of Computing & Information Sciences Vol. 7, No. 1, January 2009 On-Line.*

[2] S. Bird, M. Liberman, A formal framework for linguistic annotation, *in Speech Communication, Volume 33, Number 1, January 2001, pp. 23 -60(38) http://dx.doi.org/10.1016/S0167- 6393(00)00068-6*

[3] C. Roussey, S. Calabretto, An experiment using Conceptual Graph Structure for a Mu ltilingual Information System, *in the 13th International Conference on Conceptual Structures*, ICCS'2005

[4] F. Dau, M-L Mugnier, G. Stu mme, Conceptual Structures: Co mmon Semant ics for Sharing Knowledge, *13th International Conference on Conceptual Structures*, ICCS 2005.

[5] A. Setzer, R. Gaizauskas, TimeM L: Robust specification of event and temporal expressions in text. In *The second international conference on language resources and evaluation*, 2000.

[6] Y. Pauline Jeba, Mrs. P. Rebecca Sandra, "A Survey On Annotating Search Results Fro m Web Databases", *International Journal Of Research In Computer Applications And Robotics, Vol -1, Issue-9, 2013.*

Books:

[7] M. Islam Chisty, An Introduction to Java Annotations,2005at http://www.developer.com/java/other/article.php/109_36_3556176_1

[8] Dimitrios -Emmanuel Spanos, Periklis Stavrou a and Nikolas Mitrou a, "Bringing Relational Databases into the Semantic Web: A Survey", 0000-0000/0-1900/ $00.00c 0 – IOS Press and the authors. All rights reserved.

Theses:

[9] K. Khelif, R. Dieng-Kuntz, P. Barbry, An Ontology-based Approach to Support Text Mining and Information Retrieval in the Bio logical Do main, in J. UCS 13(12), pp. 1881-1907, 2007.

Proceedings Papers:

[10] L. Denoue, L. Vignollet, An annotation tool for Web browsers and its applications to informat ion retrieval, *in Proceedings of RIA* , 2000.

[11] P. Muller, X. Tannier, Annotating and measuring temporal relations in texts. *In Proceedings of Coling, volume I. Genève*, Association for Co mputational Linguistics, 2004.

[12] J. Kahan, M-R. Koivunen, Annotea: an open RDF infrastructure for shared Web annotations. *Proceedings of the 10th international conference on World Wide Web*, 2001.