International Journal of Research in Advent Technology, Vol.2, No.5, May 2014
E-ISSN: 2321-9637

Data Unit Level Annotation for Search Results Using Improved Alignment Algorithm

B.Jagadishkumar[1], C.Ramathilagam[2] and M.L.Valarmathi[3]
Department of Computer Science and Engineering [1, 2], Associate Professor[3]
Adithya Institute of Technology[1,2], Government College of Technology[3]
Coimbatore [1, 2, 3]
jagan.balkrish@gmail.com[1],,jkk_ramthi@yahoo.co.in[2], ml_valarmathi@rediffmail.com[3]

Abstract-The Semantic Web databases accessed through html data units are machine process able, which is fundamental for many usages such as profoundweb data collection and online evaluation shopping; they should be taken out and assigned semantic labels. In the proposed automatic annotation approach, it orders the data units on anoutcome page to various groups and the data in the same group forms. The searches results return from the databases is generally encoded into the result pages and are vibrant for the human browsing. The programmed have the meaning of same. Then, for every group, annotations are performed with certain features or attributes and collect the multiple annotations to predict a final annotation label for the group. Each site is built with an annotation wrapper or generalized class and this wrapper will annotate new result pages from the same database. The algorithm and results ensure that the proposed work gives better search response from the browsers.

Index Terms- Annotation, Data units, Semantic labels, Annotation Wrapper.

1. INTRODUCTION

Annotation-a note by way of explanation or comment added to a text or diagram.The annotation for the web database is said to be as Web Database Annotation. [11]. It gives the browser, an efficient and meaningful search results. This process of annotation is building a Semantic Web or Ontology.The World Wide Web has become one of the largest public information sources. Search engines have become the most useful tools to search the World Wide Web. Retrieval model of search engine is mainly based on looking whether keywords in a user query match the content of web documents. The search engine may omit other documents referred to the same semantic information if these documents have not the same keywords of the query. This type of annotation works on ranking algorithm and it fails in Disambigiousqueries [17].

Architecture based on software design recovery for applying the rules to mark up and extract identified instances in a document set. This Cerno, frame work purely for text based retrieval. Semantic Annotations Capture the input, Capture the output, Measure the similarity between input and output. The semantic annotation presents, a one hand model of semantic annotations for describing the Web services .The Web services uses the standards like UDDI, SOAP, and WSDL which describes only the syntactic interface of Web services. The semantic annotation lacks in OWL-Web services.

The proposed methodology considers how to automatically assign labels to the data units within the SRRs returned from WDBs. The data annotation problem and proposed a multi annotator approach [1][2][4][6].To automatically constructing an annotation wrapper for annotating the search result records retrieved from any given web database. The existing approaches just assign labels to each of the HTML text nodes. The data units are the real worldentity that represents particular object. Data units are differ from text nodes that are enclosed or embedded with pair of HTML tags.

This work is studied with the automatic data alignment problem. The Accurate alignment is critical to achieving holistic and accurate annotation. Our method is aclustering based shifting method utilizing richer yet automaticallyobtainable features.This method is capable ofhandling a variety of relationships between HTMLtextnodes and data units,including one-to-one, one-to-many,many-to-one, and one-to-nothing.

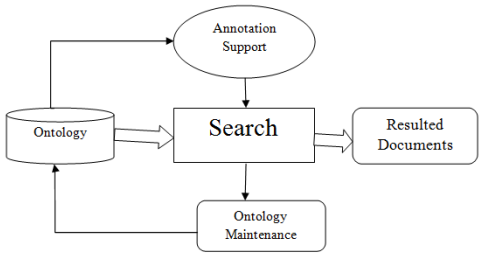

Fig: 1.System Architecture

The precision and recall measures from information retrieval to evaluate the performance of our methods [25].For alignment, the precision is defined as the percentage of thecorrectly aligned data

units over all the aligned units by thesystem; recall is the percentage of the data units that arecorrectly aligned by the system over all manually aligneddata units by the expert.The paper is organised with the following sections: section II presents the previous works for annotation, section III presents related works and techniques used in annotation. And IV section proposed work of annotation of data units using types of annotators.

2. LITERATURE SURVEY

The ranking algorithm - to re-order the results [1]. This algorithm prepared by (i) semantic annotations of web pages, (ii) Providing Annotationssemantically of queries and (iii) the logs is prepared after query searching. The algorithm makes use of this informationto elaborate an appropriate re 0rdering. To validate this ranking approach it implemented a systemthat can apply the algorithm to a particular search engine. Evaluation results show that the number ofrelevant web resources obtained after executing a query with the algorithm is higher than the oneobtained without ranking algorithm. [16] Queries and resources are semantically identified, avoiding (1) the ambiguous information stored in simple click-through data logs and (2) avoiding the limitations or restrictions in the elaboration of the queries.

Cerno, a framework is for semantic annotation of textual documents based on a domain-specific

Model which provides semi-automatic [2]. The Cerno framework is founded on light-weight techniques and tools intended for legacy code analysis and markup. To illustratethe feasibility of this framework, we report experimental results of its application to twodifferent domains. These results suggest that light-weight semi-automatic techniques for semanticannotation is feasible, require limited human effort for adaptation to a new domain, and demonstrate markup quality comparable with state-of-the-art methods.

Ontology: The ontology term comes from philosophy, which means: “the knowledge of what is to be in oneself”. In data processing, ontology indicates a structured set of knowledge in a domain. Ontology is an explicit share specification of the various conceptualizations in a particular domain [24], [22].

Annotea: A framework sematic annotations, Based on a general-purpose resource description framework (RDF) infrastructure, it describes that Annotea is a Web-based shared annotation system[4].The joining of annotations with documents presented within the client. The work on Annotea presents the overall design of Annotea and describes some of the issues faced and it is solved. Annotations for the documents are created in Resource Description Framework and Web Ontology Language servers [21]. The server stores the annotations in an RDF database. Users can query a server to retrieve an existing annotation, post a new annotation, modify anannotation, or delete an annotation. All communication between a client and an Annotea performs a simple platform for relating annotations with web texts, without the changing of documents.

Bayesian Network and Ontology based Semantic Annotation- is a semantic annotation framework that extracts and annotates information from unstructured and ungrammatical domains. It employs ontology as well as Bayesian networks (BN) to support this activity [5]. [19] Michelson and Knoblock [5] present a semantic annotation system, called Phoebus.

3. RELATED WORKS

Web information retrieval and annotation are most research area over a decade. Many systems rely on human users to mark the desired information on sample pages and label the marked data at the sometime, and then the system can induce a series of rules (wrapper) to extract the same set of information on WebPages from the same source. This system is mostly said to be as a wrapper inducing system.These systems suffer from less performance and arenot suitable for applications that need to extractRanking algorithm, known as ‘static’ algorithms, are independent of the search queries and focus on the quality of web pages by means of their inner and outer hyperlinks.

Their metric takes into account what pages users like to visit, instead of the pages web developers like to link to. There are approaches which take advantage of traditional search engines, such as web search result clustering.

When the functionality of tagging appeared, several works started to take advantage of this information to apparently produce effective ranking when searching for web pages.

$$f_p : R_q \cup R_{dq} \rightarrow [0,1] \qquad (1)$$

$$r \rightarrow \alpha . anp(r) + (1 - \alpha) . webp(r) (2)$$

Fq will contain the web resources in RqURdq sorted by theirfinal parameter fp. The constant a can be adjusted depending on user necessities or the annotations status. The architecture of Cerno is based partly on the software design recovery process of the LS/2000 system, although in that case the documents to be analyzed were computer programs written in formal programming languages, and the markup process was aimed specifically at identifying and transforming instances of Year 2000-sensitive data fields in the programs. The Cerno adaptation and generalization of this process to arbitrary text documents includes four steps: (1) document parse, (2) recognition of basic facts, (3) their interpretation with respect to a domain semantic model, and (4) mapping of the identified information to an external database.

The two semanticannotation frameworks: Phoebus and BNOSA. Phoebus uses reference sets to store domain knowledge whereas thesame task is performed by BNOSA with the aid of ontology. Both make use of different machine learningtechniques to enhance the extraction process but apply them indifferent contexts.

$$SimC(d_1, d_2) = \frac{Vd_1 * Vd_2}{||Vd_1|| * ||Vd_2||} \quad (3)$$

Similarity Measure

It is the Cosine similarity between the term frequency vectors of d1 and d2: where Vd is the frequency vector of the terms inside data unit d,||Vd|| is the length of Vd, and the numerator is the inner product of two vectors

$$SimT(d_1, d_2) = 1 - \frac{EDT(p_1, p_2)}{PLen(p_1) + PLen(p_2)} (4)$$

Tag Path Similarity

This is the edit distance (EDT) between the tag paths of two data units. Theedit distance here refers to the number of insertionsand deletions of tags needed to transform one tagpath into the other. It can be seen that the maximumnumber of possible operations needed is the totalnumber of tags in the two tag paths. Let p1 and p2 bethe tag paths of d1 and d2, respectively, and PLen(p)denote the number of tags in tag path p, the tag pathsimilarity between d1 and d2.

4. PROPOSED METHODOLOGY

The proposed work is the process of providing the annotation for every data units of the particular object. Each of the data units represents a property of the real world entity. Data units are different from text nodes that are enclosed with pair of HTML tags. This type of annotation for data units is performed with six annotators. Each annotator provides semantic information about certain entity for the applications like online shopping, online book purchasing, search engine optimization etc.

The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. In this paper, we present an automaticannotation approach that first aligns the data units on a result page into different groups such that the data in the same group have thesamesemantic.

Alignment Algorithm:

The alignment algorithm performs operation on the basis of four stages:

Stage 1: Merge text nodes. This step detects and removes decorative tags from each Search Result Records to allow the text nodes corresponding to the same attribute to be merged into a single text node

Stage 2: Align text nodes. This step aligns text nodes into groups so that eventually each group contains the text nodes with the same concept (for atomic nodes) or the same set of concepts (for composite nodes)

Stage 3: Split (composite) text nodes. This step aims to split the "values" in composite text nodes into individual data units. This step is carried out based on the text nodes in the same group holistically. A group whose "values" need to be split is called a composite group.

Stage 4: Align data units. This step is to separate each Composite group into multiple aligned groups containing the data units of the same concept.

True annotation set: the set of terms for which the object is actually annotated in database. It represents the reference set to which our predictions can be compared; *Full annotations set*: is the set of all terms found in the annotations of all interactors extracted for the requested query *Predicted annotations set*: is the set of terms predicted by the proposed computational model for the selected object.

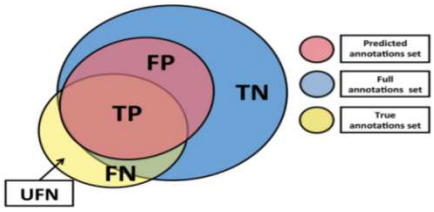

Fig: 2.Annotation Methodology Working

Common Knowledge Annotator:

A few data unit on the result page are easy to understand because of the common knowledge mutual for human beings. For example, "with stock" and "No stock" occur in many search result records from online shopping sites. So, common knowledge annotators try to make use of this state by using a few predefined familiar concepts. Each familiar concept contains a label and a set of pattern or values. The Common knowledge annotator considers both patterns and certain value sets such as the set of countries. It should be pointed out that our familiar concepts are different from the ontologies that are widely used in some works in Semantic Web. First, our familiar concepts are domain free. Following, they can be obtained from accessible information possessions with little extral effort of humans.

5. EXPERIMENTAL SETUP AND RESULTS

Protege is an open source ontology editor. Like Eclipse, Protégé is a framework for which various other projects suggest plug-in. This application is written in Java and heavily uses Swing to create the rather complex user interface. Protégé recently has over 200,000 registered users.SPARQL (pronounced "sparkle", a recursive acronym for SPARQL Protocol and RDF Query Language) is an RDF query language, that is, a query language for databases, able

to retrieve and manipulate data stored in Resource Description Framework format.

The annotation of the web databases for data units are implemented under the setup of hardware platform used during the evaluation was based on an AMD A8 Elite Quad-Core @ 2.10 GHz processor, 4 GB of DDR3 RAM the software development has been carried out in ECLIPSE on Microsoft Windows 7 Ultimate 64-bit operating system.

The annotators are mostly independent from each other since each exploits an independent feature. Based on this characteristic, we include a simple probabilistic method to combine different annotators. For a given annotator K, let M (K) be the probability that K is correct in identifying a correct label for a group of data units when K is applicable. M(K) is essentially the success rate of K. Specifically, suppose K is applicable to Vcases and among these cases J are annotated correctly, then M(K) = J/V. If a independent annotatorsK_i, i = 1,……a,identify the same label for agroup of data units, then the combined probability that atleast one of the annotators is correct is

$$1 - \prod_{i=1}^{a}(1 - M(K_i)) \qquad (6)$$

Table: 3. Performance of Data Alignment

DOMAIN	PRECISION	RECALL
AUTOMOTIVE	98.4%	98.4%
BOOK MATERIALS	98.6%	97.4%
ELECTRONICS	99%	99.1%
JOB	95.5%	100%
ENTERTAINMENT	100%	100%
OVERALL AVG	**98.3%**	**98.9%**

6. PERFORMANCE ANALYSIS

The experiments are based on five domains: book materials, entertainment, job, electronic goods, and automotives. For each WDB, its LIS is construct mechanically using WISE extractor. For each domain, WISE-Integrator is used to build the IIS mechanically. These collect WDBs are arbitrarily separated into two disjoint groups. The first group contain 22 WDBs and is used for exercise, and the second group has 90 WDBs and is used for testing. Data set DS1 is formed by obtain one sample result page from each exercise site. Two testing data sets DS2 and DS3 are generated by collecting two sample result pages from each testing site using different queries

For each result page inthis data set, the data units are manually extracted, alignedin groups, and assigned labels by a human expert. We use agenetic algorithm based method to obtain the bestcombination of feature weights and clustering threshold Tthat leads to the best performance over the training data set.

Table: 4. Performance of Annotation

DOMAIN	PRECISION	RECALL
AUTOMOTIVE	97.5%	97.7%
BOOK MATERIALS	97.3%	96.3%
ELECTRONICS	98.1%	98.1%
JOB	95.0%	100%
ENTERTAINMENT	95.9%	95.9%
OVERALL AVG	**96.7%**	**97.6%**

We adopt the precision and recall measures from information retrieval to evaluate the performance of our methods. For alignment, the precision is defined as the percentage of the correctly aligned data units over all the aligned units by the system; recall is the percentage of the data units that are correctly aligned by the system over all manually aligned data units by the expert.A result data unit is counted as “incorrect” if it is mistakenly extracted. A data unit is said to be correctly annotated if its system-assigned label has the same meaning as its manually assigned label.

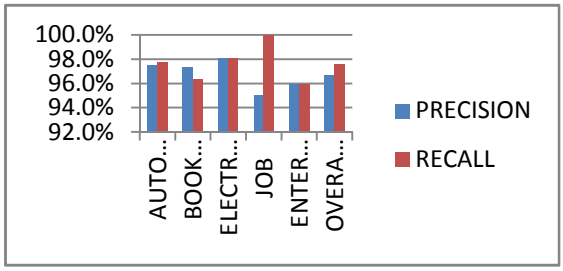

Fig: 3.Performance of Annotation

The fig: 3.analyze the performance of the annotation on precision and recall on various domains automotives, books, electronics, job, entertainment and the overall performance which have been analyzed are charted as a graph.

7. CONCLUSION

In this paper, we present an annotation for the web database underlying for the websites in various areas. The annotation for each data units provides semantic search results for the user. This method is capable of handling a variety of relationships between HTML text nodes and data units, including one-to-one, one-to-many, many-to-one, and one-to-nothing. These data units annotation carried out by six types of annotator .All type of annotator make use of one feature for annotation and outcome of this paper show that each annotator is handy and they collectively are proficient of providing improved performance in annotation.

REFERENCES

[1] Damaris Fuentes-Lorenzo, Norberto Fernandez, Jesus A. Fisteus, Luis Sanchez.” Improving large-scale search engines with semantic annotations", journal on Expert Systems with Applications from Science Direct.” Expert Systems with Applications vol.40 (2287–2296)”, 2013.

[2] NadzeyaKiyavitskaya, Nicola Zeni,James R. Cordy,LuisaMich,JohnMylopoulos.” Cerno: Light-Weight Tool Support For Semantic Annotation Of Textual Documents” a journal on

Data & Knowledge Engineering from Science Direct"Data & Knowledge Engineering vol.68 (1470–1492)",2002.

[3] HassinaNacerTalantikite , DjamilAissani, NacerBoudjlida," Semantic annotations for web services discovery and composition".a journal on Computer Standards & Interfaces from Science Direct," Computer Standards & Interfaces vol.31 (1108–1117)" ,2009.

[4] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, R.R. Swick." Annotea: an open RDF infrastructure for shared Web annotations",journal on computer networks from ELSEVIER," Computer Networks vol.39 (589–608)",2002.

[5] Quratulain Rajput, SajjadHaider."A comparison of ontology-based and reference-set-based semantic annotation frameworks" a journal on Computer Science from Science Direct vol. 3(1535–1540), 2011.

[6] Damaris Fuentes-Lorenzo, Norberto Fernandez, Jesus A. Fisteus, Luis Sanchez." Improving large-scale search engines with semantic annotations", journal on Expert Systems with Applications from Science Direct." Expert Systems with Applications vol.40 (2287–2296)", 2013.

[7] NadzeyaKiyavitskaya, Nicola Zeni,James R. Cordy,LuisaMich,JohnMylopoulos." Cerno: Light-Weight Tool Support For Semantic Annotation Of Textual Documents" a journal on Data & Knowledge Engineering from Science Direct"Data & Knowledge Engineering vol.68 (1470–1492)",2002.

[8] HassinaNacerTalantikite , DjamilAissani, NacerBoudjlida," Semantic annotations for web services discovery and composition".a journal on Computer Standards & Interfaces from Science Direct," Computer Standards & Interfaces vol.31 (1108–1117)" ,2009.

[9] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, R.R. Swick." Annotea: an open RDF infrastructure for shared Web annotations",journal on computer networks from ELSEVIER," Computer Networks vol.39 (589–608)",2002.

[10] Quratulain Rajput, SajjadHaider."A comparison of ontology-based and reference-set-based semantic annotation frameworks" a journal on Computer Science from Science Direct vol. 3(1535–1540), 2011.

[11] Zaihosnita Hood, NoraidahSahari"Researchers Annotation Collections and Practices" journal on Procedia Technology from science direct vol.11 (354 – 358), 2013.

[12] HassinaNacerTalantikite, DjamilAissani, NacerBoudjlida" journal on Computer Standards & Interfaces from science direct vol.31 (1108–1117), 2009.

[13] Raphael Volz, Siegfried Handschuh Steffen Staab, LjiljanaStojanovic, NenadStojanovic,"Unveiling the hidden bride: deep annotation for mapping and migrating legacy data to the Semantic Web" journal on Web Semantics: Science, Services and Agentson the World Wide Web vol.1 (187–206), 2004.

[14] Alexandre R.J. François and Ram Nevatia"Framework for Representing and Annotating Video Events"IEEE TRANSACTIONS ON MULTIMEDIA vol.5 (76-86), 2005.

[15] Yi Yang, Fei Wu, FeipingNie, Heng Tao Shen, YuetingZhuang, and Alexander G. Hauptmann,"Web and Personal Image Annotation by Mining Label Correlation With Relaxed Visual Graph Embedding",IEEE transactions on image processing, vol. 21, no. 3, (1339-1351),2012.

[16] Shenghua Bao1, Xiaoyuan Wu1, Ben Fei, Guirong Xue, Zhong Su, and Yong Yu,"Optimizing Web Search Using Social Annotations",the International World Wide Web Conference Committee (IW3C2) (501-510),2007.

[17] Hansen, M. H., & Shriver, E. Using navigation data to improve IR functions inthe context of web search. In Proceedings of the Tenth International Conference onInformation and Knowledge Management, (135–142), 2001.

[18] M. Michelson and C.A. Knoblock, "Semantic annotation of unstructured and ungrammatical text," In Proceedings of the 19th International Joint Conferenceon Artificial Intelligence, (1091-1098), 2005.

[19] B. Yildiz and S. Miksch, "ontoX - A method for ontology-driven information extraction," In Proceedings of the International Conference of ComputationalScience and its Application, vol. 4707, (660-673), 2007.

[20] L. Reev, H. Han, Survey of semantic annotation platforms, in: SAC'05: Proceedings of the 2005 ACM Symposium on Applied Computing, ACM Press,New York, NY, USA,(1634–1638),2005.