International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 12, December 2014 **ISSN 2319 - 4847**

GLOSSING THE INFORMATION FROM DISTRIBUTED DATABASES

E.Arun kumar[*1], S.Swapna[*2]

M.Tech Scholar, Dept of CSE, Aurora's Technological and Research Institute, Uppal, Telangana State, India,

Assistant Professor, Dept of CSE, Aurora's Technological and Research Institute, Uppal, Telangana State, India

ABSTRACT

Internet provides huge amount of useful information which is align into a format for users. Here we observe the difficulty for extraction of relevant data from different sources. Relevant data transform into structured format. Structured format contains only necessary information. The motivation behind in the system provides the compressed results which are meaningful based on concept and category. Different applications store the information in huge databases. Users are access the information from web databases based on concept wise. Single dimension concept based results are not meaningful. Meaningless records are aligning into web interfaces. In this paper we propose to extract the records with two dimensions. Those dimensions are concept and category. Using these two dimensions we organize the records into a structured format and provide the meaningful results to the users. Compare to concept we get the better results with concept and category dimensions.

KEYWORDS:- Distributed databases, e-commerce, digital libraries, concept and category based results.

1.INTRODUCTION

Extract the records from different web databases in different search engines. All records data units are encode return results display into web pages. Result page contains multiple search result records information. For example, book comparison shopping system collect the multiple records information from different book sites. Identify the labels next align the records. In previous applications human efforts in alignment of annotated data units manually. It's very complex approach. Now we move to automatic data unit alignment process. Automatic data unit alignment approach performs the operations like retrieves the records display into result pages from different number of web databases. These records extraction process we done based on concept. We got less meaningful records information. In this paper we propose the concept and category approach for retrieve the suitable or desirable records information. All records data units are align based on label wise automatically. It does provide the more meaningful results compare to all previous approaches.

2.RELATED WORK

Here we start the survey related to information extraction systems. Previously all data records are align into a DOM tree pattern. This is one of the structure based data records and data units alignment approach. Here we followed previous manual process. This is one of the limitation in DOM tree pattern approach. Next vision based approach design the two dimensions for alignment of data units. Those dimensions are spatial and temporal. These two dimensions are not work parallel. Finally this approach produces the incomplete structure for alignment of data units. Next to alignment of multi data records we introduce the ontology technology. It is work for multiple domains. All number of data units are alignment based on domain wise in the form semantic manner. Finally produce the records output based on domain dependent. This is manual process. Next here we design the new system that is called XML enabled wrapper construction system. Its process internal web pages records. Define the tags information classify the features. All features are align create the one interface. Consider the interface alignment of all records into pattern or format.

3.PROPOSED METHODOLOGY

We propose the concept and category clustering approach to alignment of data units in to different groups. In a group same semantic data units are present. DOM, Vide, ontology and xml enabled wrapper construction approaches consider only limited features in alignment of data units information. In this paper we consider more features like presentation, font styles, concept and category features also.

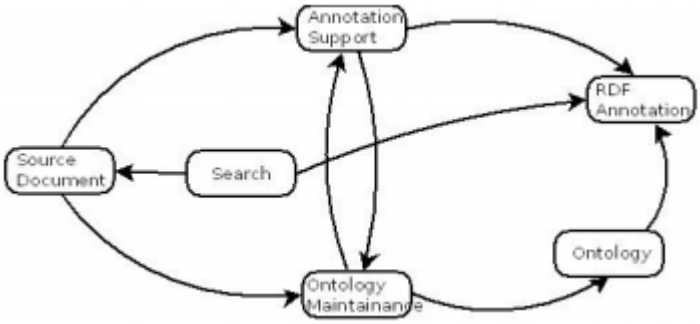

Fig3: proposed architecture

Automatic concept and category annotation solution consist of different phases. Those phases are 1. Text extraction
2. Assigning data units into a table: alignment phase, annotation phase, annotation wrapper phase.
3. Category wise annotation wrappers concepts results

Text Extraction:

Extract the text content from web pages information.

Assigning data units into a table:

Organize the data units into a different groups based on concept wise. Group content contains same meaning information. After creation of different groups identifies the common features create the pattern. That pattern we call as web interface information.

Category wise annotation wrappers concepts results:

All wrapper based concepts are not relevant. Here we categorize the wrapper patterns records based on category. These category results are more meaningful compare to concept based results.

4.RESULTS AND DISCUSSION

In this section we demonstrate the results related proposed system.

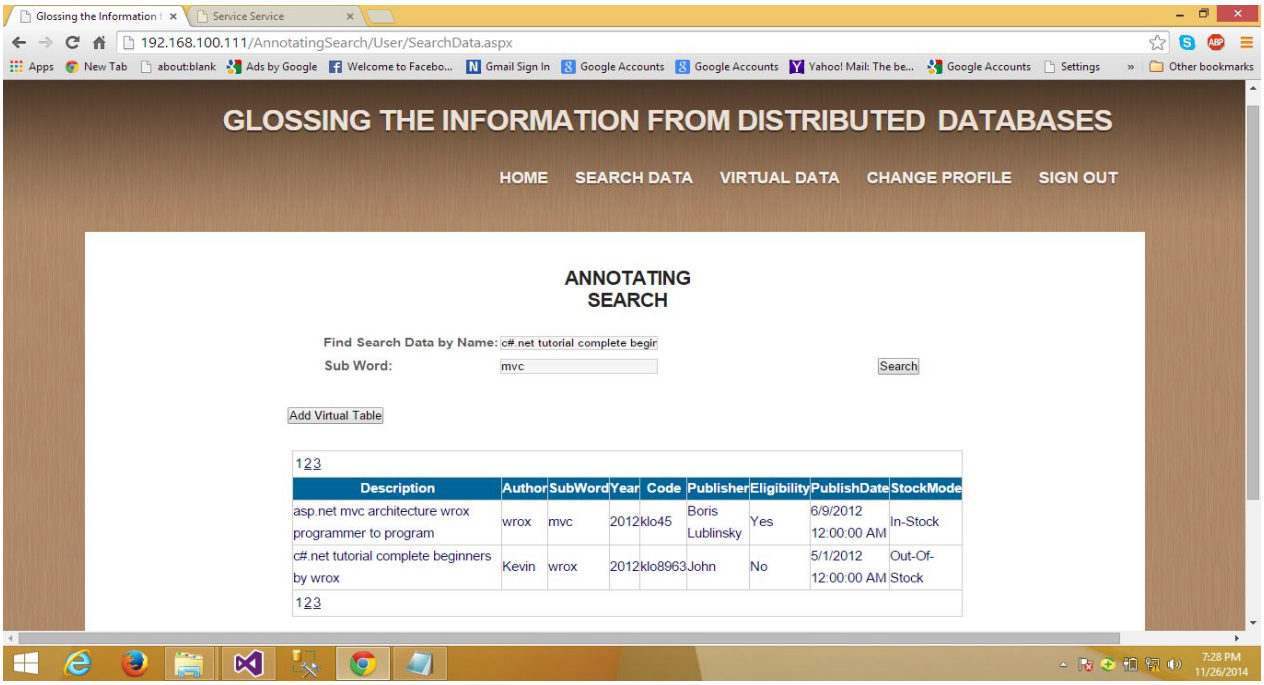

Fig1: concept, sub concept keywords based search results

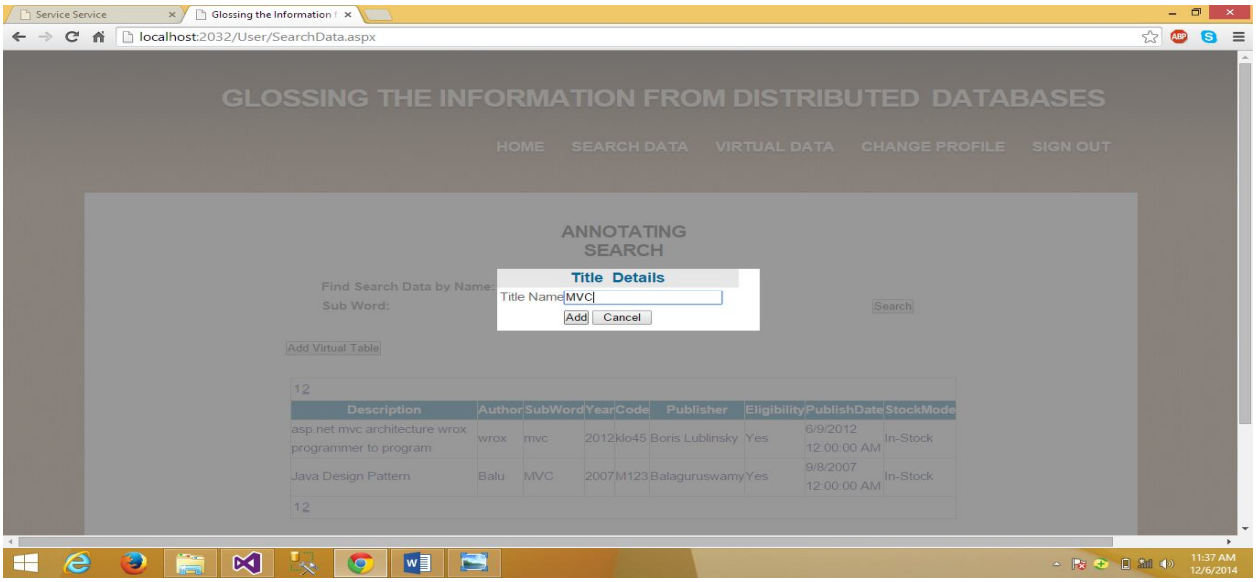

Fig2: Concept and Category keywords based results (Compressed records)

5.CONCLUSION

In this paper we design the new annotation wrapper with concept and category. This new annotation wrapper evaluate the concept based records finally produce the meaningful records information. These records are high quality and useful records to the users.

REFERENCES

[1] Annotating Search Results from Web Databases Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Member, IEEE, and Clement Yu, Senior Member, IEEE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013

[2] 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

[3] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu, "Annotating Structured Data of the Deep Web," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), 2007.

[4] STAVIES: A System for Information Extraction from Unknown Web Data Sources through Automatic Web Wrapper Generation Using Clustering Techniques Nikolaos K. Papadakis, Dimitrios Skoutas, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 12, DECEMBER 2005

[5] A Survey of Web Information Extraction Systems Chia-Hui Chang, Member, IEEE Computer Society, Mohammed Kayed, Moheb Ramzy Girgis, Member, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 10, OCTOBER 2006

[6] Wang Computer Science Department University of Science and Technology Clear Water Bay, Kowloon Hong Kong Computer Science Department University of Science and Technology Clear Water Bay, Kowloon Hong Kong

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

[8] 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.

[9] P. Chan and S. Stolfo, "Experiments on Multistrategy Learning by Meta-Learning," Proc. Second Int'l Conf. Information and Knowledge Management (CIKM), 1993.

[10] W. Bruce Croft, "Combining Approaches for Information Retrieval," Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Kluwer Academic, 2000.

[11] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc. Very Large Data Bases (VLDB) Conf., 2001.

[12] S. Dill et al., "SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation," Proc. 12th Int'l Conf. World Wide Web (WWW) Conf., 2003.

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

[14] D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y. Ng, and R. Smith, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31, no. 3, pp. 227-251, 1999.

[15] D. Freitag, "Multistrategy Learning for Information Extraction," Proc. 15th Int'l Conf. Machine Learning (ICML), 1998.
[16] D. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, 1989.