Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 9, September -2017

A Concept-Based Approach for Entity Matching

Ms. Karishma Tiware[1], Prof. Vijay Shelake[2]

[1]Department of Computer Engineering, ARMIET
[2]Department of Information Technology, YTCEM

Abstract — *One of the central problems of database integration is entity matching, that is, the identification of similar data elements in two or more databases or other data sources. This approach considers the correspondences across the attributes of various databases. Moreover, it uses different matchers to combines multiple databases. To solve the heterogeneity matching problem, we have proposed an improved approach for entity matching to increase accuracy in matching databases.*

Keywords-*Entity; Matching; Database*

I. INTRODUCTION

The data integration includes matching data from several many sources, which are stored using various technologies and provide a unified view of the data. Data integration becomes more important in cases of merging systems of two companies or consolidating applications within one company to provide a unified view of the company's data assets [8]. The major challenge is the strategy of combining data from various often incompatible sources. Matching entities is a crucial step in Web data integration, which finds attribute correspondences between data sources. The problem is closely related to schema matching that takes two schemas as input and produces a set of attribute correspondences between them. These approaches exploit different features of schemas, including structural and linguistic features and data types, etc. to match attributes between schemas. Schema matching is inherently uncertain due to lack of complete knowledge about schemas [2][6][7].

There are several organizational levels on which the integration can be performed [8]. There are several differences between query interfaces and traditional database schemas. First, database schemas are designed internally for database developers. As a consequence, the attributes of the entities may be named in a highly inconsistent manner, imposing many difficulties in entity matching. In contrast, query interfaces are designed for normal users and are likely more meaningful and consistent. For example, labels on query forms are usually words or phrases, whereas attribute names of database schemas are often abbreviations. Second, database schema matching has mainly focused on schema matching while instance-level matching has not been done extensively. This is often due to the unavailability of data instances and the assumption that the same domains of values have been used across different schemas. Whereas in query interfaces, the user is likely to be given ranges of values to choose from and as these values are designed for human use they are also likely to be more meaningful and consistent. As data instances are pervasive, semantic heterogeneity of data instances between query interfaces has to be addressed.

II. ENTITY MATCHING REVIEW

Entity matching methods are primarily categorized by their use of schema-level or instance-level information, although many methods use both types of information. Schema-level matching methods may also use structural and constraint information such as relationship types between entity types or foreign-key dependencies between tables. The most of the previous approaches basically depend on element name more than other information, while some approach exploits some available information, DBMS catalog, data type and thesaurus, in order to improve the accuracy of the matching. Few other hybrid methods, take into account the synonyms related to the element name. Unfortunately, only these approaches are still limited, since they are dependent on comparing the element name, by using the semantic similarity measures without considering the lexical ambiguity among database schema.

III.PROPOSED SYSTEM ARCHITECTURE

In proposed system, we have proposed architecture for entity matching by using concept matching for providing high level accuracy by including all the available approach and further providing disambiguity among concepts. It includes synchronization of entity matching, concept word and semantic matching. The Figure 1 shows the proposed system architecture for entity matching [1][2][3][4]:

- **Loading process**
  The input is the original schema set from two individual database from different interfaces, and the output is a global schema set for constituting the domain uniform interface. All these approaches concentrate on finding the most plausible correspondences among schema, by using string comparing of the schema elements name (attribute) together in the first step, then filtering the match result by using other available information such as key constraints, instance data, data type and some auxiliary sources.

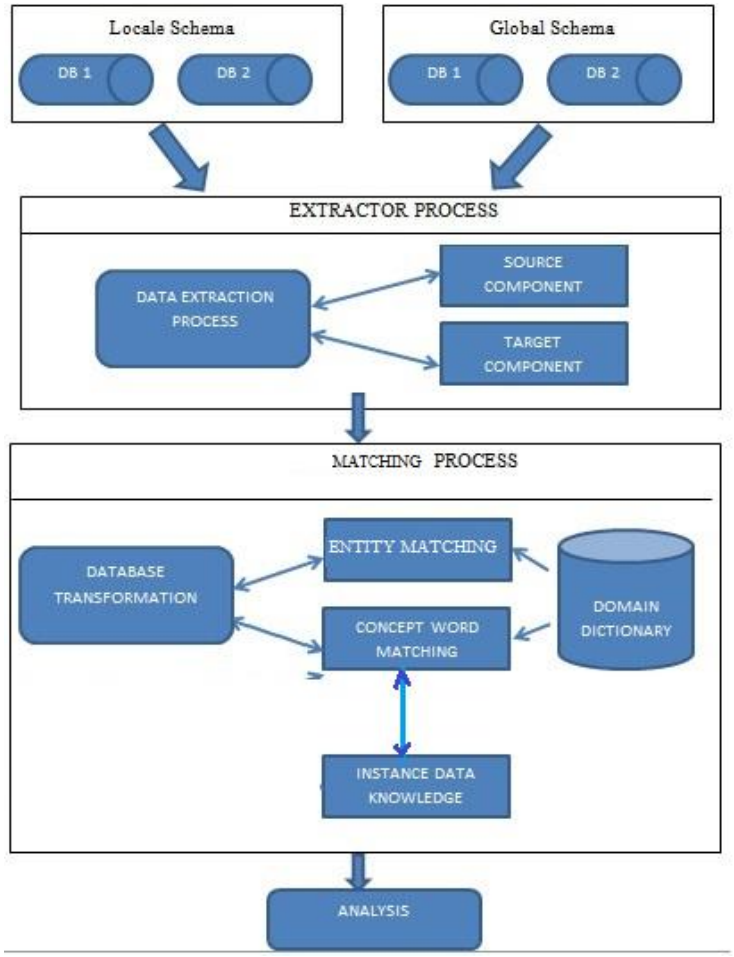

Fig 1: System Architecture

- **Extraction and Matching Process**
  Extraction and matching process includes extraction, transformation and disambiguation process. While a computer has no built-in mechanism for understand the semantics of words and symbols, a dictionary is required to help the computer to determine meaning of words and symbols. In this work, we use domain dictionary, to provide the proposed disambiguation with knowledge which used to recognize lexical ambiguity and decide the most possible related sense of the word.

- **Concepts Words**
  The concept-word refers to some words that can be used as the attribute name alone and can represent essential meaning of the attribute.

- **Semantic Matching**
  The semantic matching refers to the attributes with the same meaning but different attribute names in different interfaces.

IV.DISCUSSION

In this paper, the system architecture has been proposed for entity matching. Element matching is used along with concept level matching and hierarchical dictionary is used to enhance accuracy level and matching attributes under different entities. In addition, the model will be tested to identify results for the process of entity matching.

V.CONCLUSION AND FUTURE WORK

In this paper, we have described a general framework for database integration of semantically different databases and presented compelling direction of this framework. Further it is planned to involve more identical database with higher data integration support to provide better accuracy and runtime.

REFERENCES

[1] Li-Jun Chen, "Deep Web Schema Matching Based on Concept-word and Semantic-heterogeneous Model," 2016 3rd International Conference on Information Science and Control Engineering.

[2] W. Liu, X. F. Meng, and W. Y. Meng, "A survey of Deep Web data integration," Chinese Journal of Computers, vol. 3, no. 9, pp. 1475-1489, 2007.

[3] J. Hong, Z. T. He, and D. Bell, "An evidential approach to query interface matching on the Deep Web," Information Systems, vol. 35,no. 2, pp. 140-148, 2010.

[4] G.A. Miller, WordNet: a lexical databases for English, Communications of ACM 38 (11) (1995) 39–41.

[5] Y. Karasneh, et al., "A model for matching and integrating heterogeneous relational biomedical databases schemas," in Proceedings of the International Database Engineering &Applications Symposium, 2009, pp. 242-250.

[6] Bilke, A. and Naumann, F., Schema Matching using Duplicates, Proceedings of the Twenty-first International Conference on Data Engineering, 2005, pp. 69-80.

[7] Dhamankar, R., Lee, Y., Doan, A., Halevy, A., and Domingos, P., iMAP: Discovering Complex Semantic Matches between Database Schemas, Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, 2004, pp. 383-394.

[8] http://www.dataintegration.info/data-integration

