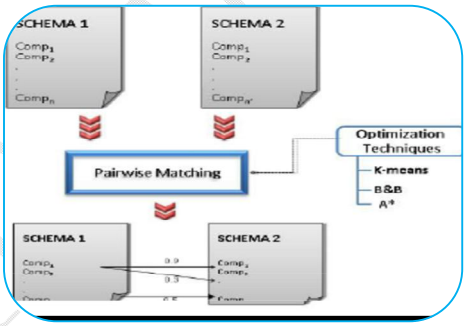
REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
UGC APPROVED JOURNAL NO. 48514
VOLUME - 8 | ISSUE - 9 | JUNE - 2019

"THE STUDY ON OVERVIEW OF DIFFERENT SCHEMA MATCHING TECHNIQUES"

Shilpa Deshmukh , Dr. Pravin Karde and Dr. Vilas Thakare

ABSTRACT:



Schema matching is a primary challenge in many database applications such as data integration, E-business, data warehousing, and semantic query processing. The main aim of the schema matching process is to identify the correlation between schema which helps later in the data integration process. The main issue concern of schema matching is how to support the merging decision by providing the correspondence between attributes through syntactic and semantic heterogeneous in data sources. We present a taxonomy that covers many of existing approaches, and we describe the approaches in some detail. In particular, we distinguish between schema-level and instance-level, element-level and structure-level, and language-based and constraint-based matchers. We intend our taxonomy and review of past work to be useful when comparing different approaches to schema matching.

KEYWORDS: *Data Integration, individual matcher, Schema Matching, semantic, structure level matching.*

INTRODUCTION

In recent times integration and handling of a huge quantity of data has been tremendously simplified because of the development in information technology. Numerous solutions have been put forward to integrate data from various heterogeneous sources to form a integrated global view. This, procedure which aims to represent data in one single view and enable the process of communicating with the data for being app eared as one single information system[1] is called as data integration. Though, it is very challenging to assimilate and manage data from various sources that are developing individually. This is because of the fact that there are various representations of these sources, and those might not be built to implement the same principles or have similar semantic concepts to be fully used [2]. Moreover, there might be various terminologies used to define and store information which may inversely affect the process of integrating the data [3]. Organizations with various databases try to integrate developed heterogeneous data sources and each database may comprises of huge number of tables that incorporate different attributes. The heterogeneity in the data sources can lead to growth of the complexity of handling these data, which result in the need for data integration [4]. Recognizing the clash of (syntax and semantic heterogeneity) between schemas is a significant challenge through data integration process. Hence, schema matching has been

projected to manage the process of noticing the correspondence between schema and resolve clash when happened.

Simplest form of schema matching comprises of recognizing two elements from two different schemas as semantically equivalent, or matched. A primary task in the handling of schema information is Matching, which takes two schemas as input and gives a mapping between elements of the two schemas that correspond semantically to each other as output [5,6,7,8,9,10]. Matching is a central basic process in several applications, such as web-oriented data integration, electronic commerce, schema integration, schema evolution and migration, application evolution, data warehousing, database design, web site creation and management, and component-based development. [11]

Moreover, as organizations turn out to be clever to manage databases and applications with more complexity, their schemas become bigger, growing the number of matches to be achieved. The amount of work is at least linear in the number of matches to be performed, may be worse than linear if one needs to evaluate each matching the context of other possible matches of the same elements. A faster and effort less integration approach is required. This entails automated support for schema matching.

Fortunately, there is a plenty of prior research on schema matching built in the framework of schema translation and integration, knowledge representation, machine learning, and information retrieval. The main goals of this study are to study these previous approaches and to present a taxonomy which describes their common characteristics. We assume the study to be obliging in designing new methods and in selecting approach from a library of approaches to use.

This paper is organized as : ii) use of schema matching in application domain, iii) Criteria for comparison, iv) Classification of approaches of schema matching ,iv) Discussion v) Conclusion.

1. Domain of Application

To understand the need and importance of schema matching study , we should know possible application where schema matching is an integral part of .We tried to brief some of them

i. Schema Integration:

It is an activity which offers a unified representation of multiple data sources. The fundamental challenge in schema integration are: schema matching [1], i.e. the identification of correspondences, or mappings, between schema objects, and schema merging [2], i.e. the creation of a unified schema based on the identified mappings[12] .As the schemas are individualistically developed, mostly they have unlike structure and terminology. This can apparently occur when the schemas are from different domains, such as a real estate schema and property tax schema. Though, it also occurs even if they model the same real world domain, just because they were developed by different people in different real-world contexts. Hence, a first step in schemas integration is to recognize and illustrate these interschema relationships. This is a process of schema matching. Once they are recognized, matching elements can be unified under a comprehensible, integrated schema or view. In this process of integration, or occasionallymay be as a distinctphase, programs or queries are designed which permit translation of data from the original schemas into the integrated presentation. So basically schemaintegration is a process of integration of independently developed schema with a given conceptual schema. This process needsreconciliationof the structure and terminology of the differenttwo schemas, which comprises schema matching.

ii) Data Warehouses:

A data warehouse is a decision support database that is taken out from a set of data sources. This process of extraction requires converting data from the source format into the warehouse format. As shown in [13], the process of matching is useful for designing transformations. An approach for making suitable changes is to begin by searching elements of the source that are also present in the warehouse for a certain data source . This isprocess of matching. Once an initial mapping is done the

data warehouse designer requires examining the detailed semantics of each element and generating transformations that resolve those semantics with those of the goal.

iii) E-commere

The fundamental goal of data integration for e-commerce is to share data. E-commerce has directed to a new inspiration for schema matching is transformation of message. Business partnersoften interchange messages that definedealings for their ongoing business. Each business organization uses their own format for the message; hence it may vary in structures.

Message schemas may have different structure. Application developers supposed to translate messages into the formats prescribed by various businessorganizations to empower systems to exchange messages. Message translation is translations between different message schemas. These Message schemas may include names, data types, and ranges of allowable values which may be completely unlike in each message schema. Schema matching methods requireidentifying such semantic differences to match such schema elements appropriately.

iv) Sementic Web

Identifying the vision of a semantic web is consistent data integration. Semantic query processing occurs in a run-time scenario where the results of a query are mentioned. This generally happens in case of 'deep web' or dynamic web which can accessed only through web forms for example flight tickets availability as per requirement or books availability on particular subject in a library. These databasescan be accessed only through filling in and submitting a form on the web. As perprecise views retrievingprimary web databases, and are related to schema matching, web forms can be abstracted because information needed by a form or query interfaces on the web can be assumed as a form of schema for the database (Halevy 2005). Web search engines generally are not able to access such information and so is likely to be unavailable.

3. COMPARATIVE MEASURES

To compare the evaluations of schema matching approaches we consider criteria from four different areas:

Input: What type of input data has been accepted (information of schema, data instances, dictionaries etc.)? To get more accurate results, more detailed and supportive information should be used.

Output: Type of information has been comprised in the match outcome.Correct evaluation of result.Minimum information as output will decrease the chances of errors but will increase post-processing effort.

Quality measures: type of metrics have been chosen to measure the correctness and completeness of the match result?

Effort: how the manual effort is minimized and how it is counted. Type of manual effort measured may be pre-match or post-match.

2. Classification of approaches of schema matching

There are large number of schema matching techniques designed to recognize the match among the database tables. [14] and [15] studied classified and surveyed number of approaches.Schema Matching approaches broadly classified into two types: individual matcher and combining matchers as showed in Fig. 1

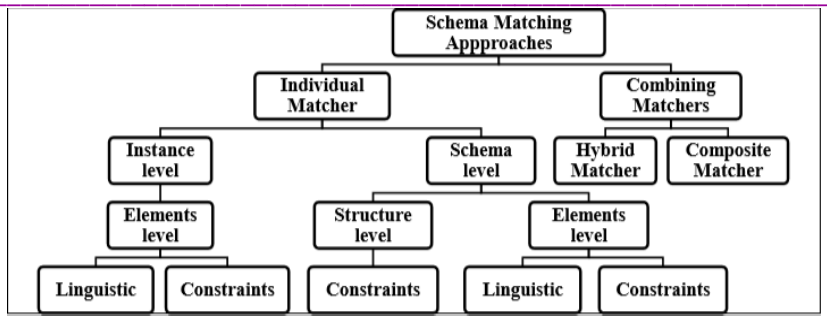


Fig 1: Classification of schema matching techniques

To compute the mapping among instances,individual matchers considers only one single parameter. It only focuses on offered schema metadata for

A. Individual Matcher

Individual matchers concentrate on the available schema metadata (metadata) in terms of integrity constraints, attributes names, descriptions, and schema structures with disregard to the lowest level of information (instance) [14]. Schema information is commonly used to accomplish the matching between simple databases, and it is very helpfulwhen instance level data is unavailable[16]. Contrary to this, combining matchers either includesmultiple criteria (eg.name and type equality) to design hybrid matcher or merge multiple match results from many matchingapproaches to form a compound match. Further all types of the schema matching techniques are discussed:

1. Schema Level Matching

Schema-level matching does not consider instance data rather only focus on schema information. The existing information such as name, description , datatype, constraint and structure of schema is used by this type of matching to detect the matching among two attributes of the schemas of the database. Usually, multiple candidate match could be produced for individual candidate, with assessed degree of resemblance in the scale of 0 to 1. The match is considered as more similar if the degree of similarity is closer to one.Element level and structure level are the two levels under schema level matching are used to define the equivalence between attributes.

a. Element Level Matching:

Element level matching triesto engage the elements of thesource schema to decide the matching elements of the input target schema. Usually, the schema elements can be employed at the optimum level also called as atomic level, such as attributes in an XML schema or columns in a relational schema[11].Element level matching also focuses on exploiting two levels that are linguistic matcher and constraint matchers.

i. Linguistic Matcher

It includes the existing linguistic information of the database schemas such as attributes names and descriptions of the attributes to decide the match between the schemas [17]. This method is frequently used for databases in unified environment, in which meaning of attribute defines the resemblance between attribute name. Linguistic matcher is useful in semi structured databases also withwell- defined schema descriptions also.

ii. Constraint Matcher

Constraints are employed database schemas for characterizing the data types, the range of values, the exclusivity, the types of relationships and cardinalities[18]. Generally, if both schemas have enough amount of constraint information, it will be used by matcher to find out match between schema

and schema attributes. For example, match score can be defined based on the similarity of datatypes or domains. Moreover many key characteristics can be included to to measure the similarity score, primary key and foreign key as well[18]. Sometimes constraint information might cause fallacious match because of comparable constraints between attributes in schemas. Though discovering constraints information always helpful to decrease the number of match which might be blended with some other matcher [2][11][19].

b. Structure Level Matching

In structure level matching structural information about the database schemas is used to discover the match between schemas. It focuses on the structures and the constraints information about the aimed schemas to derive the match between the attributes[20]. Depending on the integrity of the structural information and appropriate accuracy, there are many possible combination of various attributes in a structure. Ideally all the attributes of the source and target schemas should match with each other. Though in some cases, partial match is allowed when there is a comparison between sub-schemas.(Notice the example given in Table 1, where partial match occurred between Account Owner and Customer schemas). Rather Constraint-based matcher can be used in this level, accomplishing the constraints information such as data types, value ranges, nullability, and referential integrity (foreign keys) [2], [21], [22], [23],[24].

2. Instance Level Matching

Generally with semi-structured databases, information might not be available or inadequate for required schema matching result[2][22][20][25]. Hence exploitation of schema information is not possible all the time to accomplish a correct match among schemas.For such circumstances instances are used in place of source for deciding respective attributes. Instance level methods use the available instance as a source to find the correlation between schema attributes. Instance level data is powerful substitute source providing toward correct matching because of its treasured contents and the meaning of schema attributes.

B. Combining Matchers

After studying and evaluating various schema matcher, it has been noticed every technique has its merits and demerits. There is hardly single approach which fits for all cases and provide a reliable match. For this reason , need of combining two matchers have been aroused. Combiningmatching approaches also raises the chance to assess them concurrently or in a precise order.

1. Hybrid Matcher

Hybrid matcher combines multiple matching approaches to find match which depends on number of criteria and origin of the information. This comprises name matching and thesauri joint with data types to give exact matching results. This approach maintains high performance if compared with separated individual matcher. Single match candidate matches only one of several criteria which can be refined out early and hence efficiency may be improved complex matches requiring the joint consideration of multiple criteria can be solved[11]. Because of these two reasons efficiency can be made better.Structure-level matchingalso aidsof being combined with other approachessuch as name matching. One way to combine structure- withelement-level matching is to use one algorithm to generate apartial mapping and the other to complete the mapping[26].

2. Composite Approaches

Composite matcher first aims to conduct the independent match on database schemas using different techniques and then combine the results of both. This permits to implement the selection of the most appropriate matcher.Composite matcher is more flexible as compare to hybrid matcher as it employs the application domain and input schemas information, whereas the other methods can be used for structured versus semi-structured schemas [11], [22], [27]. Moreover, a compositematcher

generally permits a flexible organization of matchers in a such way that they are either implemented simultaneously or sequentially. In thefollowinginstance, the match result of a first matcher is expended and protracted by a second matcher to attain a repetitive improvementof the match result.

Matchers are selected, their execution order and combination of independently defined match results are defined either automatically by the implementation of Match itself or its clients or manually by a human user.Though an automatic approach decreases the number of user interactions, it is not easy to attain a common solution which can be revised to different application domains. Instead, a user can select the matchers to be implemented with their execution order and the way of combining their results. This type of approach is easier to implement and gives better control to the user.

4. DISCUSSION

From aboveclassification and its applications it can summarised that schema matching among heterogeneous databases is a critical task. Most of the techniques use metadata information to deal with this issue[9][11],though it is not enough. So we are listing out some of the areas which need attention. Because of fast growth in the data volumes, Big data is giving lot of oppurtunities to the researcher. A hot research area which should be exploited in big data is schema matching where tens or hundreds of millions of records and analyzing the sample might lead to an exhaustive process that consumes a significant amount of time. Another issue is incomplete databases.These incomplete and inaccurate data have a negative influence on the consistency of the matching results. Hence, many applicationsclaimed that the results extracted from sampling include inaccurate, or incomplete data should not be trusted [44], [46]. Next area is of uncertain databases, the values are not distinct and fluctuate in a range of values [45]. Data uncertainty can also have a negative influence on the matching process and the accuracy. Thus, it would be attainable to implement directly the conventional instance-based schema matching technique on undefined databases as it might sustaingreater processing cost and negotiating the match quality.

Various techniques have been offered, implemented, several schemes for accurate determination of correspondence between attributes of schemas. From the literature, it can be epitomized that four main schemes which can discover the contents of the database (instances) to detect the correspondence between attributes that directs to schema matching [31], [32]. These schemes are neural network, machine learning, information theoretic discrepancy and rule based. Our research suggests to discover more in above mentioned areas.

5. CONCLUSION

Schema matching is a primary problem in several database application fields, such as heterogeneous database integration, E-commerce, data warehousing, and semantic query processing. Schema matching aims at discovering the correspondences between attributes of database schemas. This paper provides a comprehensive classification of schema matching approaches schema matching. In particular, we distinguished between schema level and instance-level, element level, and structure level, and linguistics and constraint matchers, and discussed the combination of multiple matchers (hybrid and composite matcher). We used the taxonomy to characterize and compare a variety of previous match implementations.

REFERENCES

[1]M. Lenzerini, "*Data integration: A theoretical perspective,"* Proceedings of the 21st ACM SIGMOD-SIGACT-SIGARTsymposium on principles of database systems. ACM, pp. 233-246, 2002.

[2]M.A. Osama, I. Hamidah, and A. S. Lilly, "*An approach for instance based schema matching with Google similarity and regular expression,*" The Int. Arab J. of Info. Tech. Pp. 755- 763, 2017.

[3]R.Gligorov, W. Ten Kate, Z. Aleksovski and F. Van Harmelen, *"Using Google distance to weight approximate ontology matches,"*Proceedings of the 16th international Conference on World Wide Web. ACM, pp. 767-776, 2007.

[4]P. P. A. L. Leme, M. A. Casanova, K. K. Breitman, and L. A. Furtado, *"Instance-Based OWL schema matching,"* Proceedings of the 11th International Conference, ICEIS 2009, pp. 14- 26, 2009 .
[5]Li W, Clifton C (1994) *Semanticintegration in heterogeneous databases using neural networks.* In: Proc20th Int Conf On Very Large Data Bases, pp. 1–12
[6]Miller R,YEIoannidis, RamakrishnanR(1994*) Schema and practice.* Inf Syst 19(1):3–31
[7]Milo T, Zohar S (1998*) Using schema matching to simplify heterogeneous data translation*. In: Proc24th Int Conf On Very Large Data Bases, pp. 122–133
[8]Palopoli L, Sacca D, Ursino D (1998) *Semi-automatic, semanticdiscovery of properties from database schemas*. In: Proc Int. Database Engineering and Applications Symp. (IDEAS), IEEE
Comput, pp. 244–253
[9]Mitra P, Wiederhold G, Jannink J (1999*) Semiautomaticintegration of knowledge sources*. In: Procof Fusion '99, Sunnyvale, USA,
[10]Doan AH, Domingos P, LevyA (2000) *Learning source descriptions for data integration*. In: ProcWebDBWorkshop, pp. 81–92
[11]Erhard Rahm Philip A. Bernstein "*A survey of approaches to automatic schema matching*" December 2001, Volume 10, Issue 4, pp 334–350
[12]Matteo Magnani Nikos Rizopoulos Peter Mc.Brien Danilo Montesi " Schema Integration Based on Uncertain Semantic Mappings"International Conference on Conceptual Modeling 2005, Conceptual Modeling – pp 31-46
[13]Bernstein PA, Rahm E (2000) "*Data warehouse scenarios for model management*." In: Proc19th Int Conf On Entity-Relationship Modeling, Lecture Notes in Computer Science, vol. 1920. Springer, Berlin Heidelberg NewYork, 2000, pp. 1–15
[14]A. P. Bernstein, J. Madhavan and E. Rahm,(2011) "*Generic schema matching, ten years later,"* Proceedings of the 37th International Conference on Very Large Data Bases, 4(11), pp. 695-701.
[15]T. B. Dai, N. Koudas, D. Srivastava, K. A. Tung, and S. Venkatasubramanian(2008) "*Validating multi-column schema matchings by type,"* Proceedings of the 24th International Conference on Data Engineering, IEEE, pp. 120-129.
[16]H. D. Hong, and E. Rahm, *"COMA: a system for flexible combination of schema matching approaches,"* Proceedings of the 28 International Conference on Very Large Data Bases, VLDB Endowment, pp. 610-621, 2002.
[17] P. A. Ambrosio, E. Métais, and N. J. Meunier, *"The linguistic level: contribution for conceptual design, view integration, reuse and documentation,"* Data & Knowledge Engineering, 21(2), pp. 111-129, 1997
[18] Ali Alwan, Mogahed Alzeber, AzlinNordin, Abedallah Zaid Abualkkishik*" A Survey of Schema Matching Research using Database Schemas and Instances"* International Journal of Advanced Computer and Applications Vol 8. No.10, 2017, p no 102-109
[19] K.S.Zaiss*"Instance-based ontology matching and the evaluation of matching systems*," Unpublished doctoral Dissertation. University of Dusseldorf, Germany, 2010.
[20] S. Jain and S. Tanwani, "*Schema matching technique for a heterogeneous web database,"* Proceedings of the 4th International Conference on the Reliability, Infocom Technologies and Optimization (ICRITO) Trends and Future Directions). IEEE, pp. 1-6, 2015
[21] H. Do, *"Schema matching and mapping-based data integration: architecture, approaches, and evaluation,"* Saarbrucken, German: VDM Verlag, 2007.
[22] O. A. Mahdi, I. Hamidah., and S. A. Lilly, "Instance based matching using regular expression," Procedia Computer Science, 10, pp. 688-695, 2012.
[23] K. S. Zaiss, "Instance-based ontology matching and the evaluation of matching systems," Unpublished doctoral Dissertation. University of Dusseldorf, Germany, 2010.
[24]S. Anam, S. Y. Kim, H. B. Kang, Q. Liu, "Review of ontology matching approaches and challenges," International Journal of Computer Science and Network Solutions, 3(3), pp. 1-27, 2015.
[25]G. M. De Carvalho, H. A. Laender, A. M. GonçAlves, and S. A. Da Silva, "An evolutionary approach to complex schema matching," Information Systems, 38(3), pp. 302-316, 2013.

[26]Been-CfflanChien,Shiang-Yi He , "*A hybrid approach for automatic schema matching*" Conference Paper: Machine Learning and
Cybernetics (ICMLC), 2010 International Conference on, Volume: 6
[27] Y. Gozudeli, H. Karacan, O. Yildiz, M. Baker, A.Minnet, M. Kalenderand M. Akcayol, "*A new method based on Tree simplification and schema matching for automatic web result extraction and matching,*" Proceedings of the International MultiConference of Engineers and Computer Scientists. Hong Kong, China, IMECS, pp. 1-5, 2015.