Holistic Schema Matching

Erhard Rahm, Eric Peukert

Synonym

Collective Schema Matching

Definition

Holistic schema matching aims at identifying semantically corresponding elements in multiple schemas, e.g., database schemas, web forms or ontologies. The corresponding elements from N (>2) sources are typically grouped into clusters with up to N members. Holistic schema matching is usually applied when multiple schemas need to be combined within an integrated schema or ontology.

Overview

Holistic schema matching aims at identifying semantically corresponding elements in multiple (>2) schemas, such as database schemas, web forms or ontologies. It is to be contrasted with the traditional pairwise schema matching (Rahm and Bernstein 2001, Euzenat and Shvaiko 2013) between two input schemas only that determines a so-called mapping consisting of a set of correspondences, i.e., pairs of elements of the input schemas (table attributes, ontology concepts) that match with each other. Holistic schema matching is applied to more than two input schemas so that matching schema elements are clustered or grouped together.

This form of schema matching is needed to support data integration for more than two data sources and is thus of increasing relevance for Big Data applications. There is a large spectrum of possible use forms for holistic schema matching depending on the number and complexity of schemas. Traditional approaches for schema and ontology integration (Batini et al 1986, Raunich and Rahm 2014) are typically limited to few sources with a high degree of manual interaction. The need for automatic approaches has increased by the availability of many web data sources and ontologies. In domains such as the life sciences, there are numerous overlapping ontologies so that there is an increasing need to semantically integrate these ontologies into a combined ontology to consistently represent the knowledge of a domain. An example of such an integration effort is the biomedical ontology UMLS Metathesaurus (Bodenreider 2004) which currently combines more than three million concepts and more than 12 million synonyms from more than 100 biomedical ontologies and vocabularies. The integration process is highly complex, and still largely dependent on domain experts, in order to identify the synonymous concepts across all source ontologies as well as to derive a consistent ontology structure for these concepts and their relations. Related to the integration of multiple ontologies is the construction of so-called *knowledge graphs* from different web sources and ontologies (Dong et al 2014) that requires to holistically integrate both entities of one or several domains (e.g., countries, cities, presidents, movies, etc.) and schema elements such as attributes, relationships and concepts.

While the challenges for integrating very large sets of schemas are far from being solved there are initial approaches for holistic schema matching that we will discuss in this entry, in particular for the following use cases (Rahm 2016):

- Integration of several database schemas or ontologies.
- Matching and integrating multiple *web forms*, e.g., for a mediated web search over several databases.

- Matching several *web tables* of a certain domain e.g., for improved query processing.

Ideally, holistic schema matching is largely automatic with minimal user interaction and achieves a high match quality, i.e., it correctly clusters matching schema elements. Furthermore, holistic match approaches need to be efficient and scalable to many schemas/sources. It should also be easily possible to add and utilize additional schemas and deal with changes in the source schemas.

As we will see holistic schema matching still uses pairwise matching as a building block before a clustering takes place. However, scalability to many sources is not feasible if a pairwise mapping of all N input schemas is needed since this would result in a quadratic number of mappings, e.g., almost 20,000 mappings for 200 sources (Rahm 2016).

Key Research Findings

In the following, we will discuss main approaches for the mentioned use cases.

Holistic matching for schema and ontology integration

To match and integrate multiple schemas or ontologies one can in principle apply a pairwise matching and integration (or merging) multiple times in an incremental way. For instance, one can use one of the schemas as the initial integrated schema and incrementally match and merge the next source with the intermediate result until all source schemas are integrated. This has the advantage that only N-1 match and integration steps are needed for N schemas. Such binary integration strategies have already been considered in early approaches to schema integration (Batini et al. 1986). More recently it has been applied within the Porsche approach (Saleem et al 2008) to automatically merge many tree-structured XML schemas. The approach holistically clusters all matching elements in the nodes of the integrated schema. The placement of new source elements not found in the (intermediate) integrated schema is based on a simplistic heuristic only. A general problem of incremental merge approaches is that the final merge result typically depends on the order in which the input schemas are matched and merged.

A full pairwise matching has been applied in (Hu et al 2011) to match the terms of more than 4000 web-extracted ontologies. The match process using a state-of- the-art match tool took about one year on six computers showing the insufficient scalability of unrestricted pairwise matching.

A holistic matching of concepts in linked sources of the so-called Data Web has been proposed in (Gruetze et al 2012). The authors first cluster the concepts (based on keywords from their labels and descriptions) within different topical groups and then apply pairwise matching of concepts within groups to finally determine clusters of matching concepts. The approach is a good first step to automatic holistic schema matching but suffered from coverage and scalability limitations. So topical grouping was possible for less than 17% of the considered one million concepts and matching for the largest group of 5K concepts took more than 30 hours.

The matching between many schemas / ontologies can be facilitated by the re-use of previously determined mappings (links) between such sources, especially if such mappings are available in the Data Web or in repositories such as Bio-Portal and LinkLion (Nentwig et al 2014). Such a re-use of mappings has already been proposed for pairwise schema and ontology matching based on a repository of schemas and mappings (Do and Rahm 2002, Madhavan et al 2005, Rahm 2011) A simple and effective approach is based on the composition of existing mappings to quickly derive new mappings. In particular, one can derive a new mapping between schemas S1 and S2 by composing existing mappings, e.g., mappings between S1 and

S_i and between S_i and S2 for any intermediate schema S_i. Such composition approaches have been investigated in (Gross et al 2011) and were shown to be very fast and also effective. A promising strategy is to utilize a hub schema (ontology) per domain, such as UMLS in the biomedical domain, to which all other schemas are mapped. Then one can derive a mapping between any two schemas by composing their mappings with the hub schema.

Holistic matching of web forms and web tables

The holistic integration of many schemas has mainly been studied for simple schemas such as web forms and web tables typically consisting of only a few attributes. Previous work for web forms concentrated on their integration within a mediated schema while for web tables, the focus has been on the semantic annotation and matching of attributes.

The integration of *web forms* has been studied to support a meta-search across deep web sources, e.g., for comparing products from different online shops. Schema integration mainly entails grouping or clustering similar attributes from the web forms, which is simpler than matching and merging complex schemas. As a result, scalability to dozens of sources is typically feasible. Fig. 1 illustrates the main idea for three simple web forms to book flights. The attributes in the forms are first clustered and then the mediated schema is determined from the bigger clusters referring to attributes available in most sources. Queries on the mediated schema can then be answered on most sources.

Fig. 1: Integration of web forms (query interfaces) into a mediated schema for flight bookings

Proposed approaches for the integration of web forms include Wise-Integrator and MetaQuerier (He et al 2004, He and Chang 2003). Matching and clustering of attributes is mainly based on the linguistic similarity of the attribute names (labels). The approaches also observe that similarly named attributes co-occuring in the same schema (e.g., FirstName and LastName) do not match and should not be clustered together. (Das Sarma et al 2008) proposes the automatic generation of a so-called probabilistic mediated schema from N input schemas, which is in effect a ranked list of several mediated schemas.

A much larger number (thousands and millions) of sources has to be dealt with in the case of web-extracted datasets like *web tables* made available within open data repositories such as data.gov or datahub.io. The collected datasets are typically from diverse domains and initially not integrated at all. To enable their usability, e.g., for query processing or web search, it is useful to group the datasets into different domains and to semantically annotate attributes.
Several approaches have been proposed for such a semantic annotation and enrichment of attributes by linking them to concepts of knowledge graphs like Yago, DBpedia, or Probase (Limaye et al 2010, Wang et al 2012). In (Balakrishnan et al 2015) attributes are mapped to concepts of the Google Knowledge Graph (which is based on Freebase) by mapping the attribute values of web tables to entities in the knowledge graph and thus to the concepts of these entities.

The enriched attribute information as well as the attribute instances can be used to match and cluster related web tables to facilitate the combined use of their information. This has only been studied to a limited degree so far. In particular, the Infogather system (Yakout et al 2012) utilizes enriched attribute information to match web tables with each other. To limit the scope they determine topic-specific schema match graphs that only consider schemas similar to a specific query table. The match graphs help to determine matching tables upfront before query answering and to holistically utilize information from matching tables. (Eberius et al 2013) uses instance-based approaches to match the attributes of web tables considering the degree of overlap in the attribute values, but they do utilize the similarity of attribute names or their annotations.

The holistic integration of both web forms and web tables is generally only relevant for schemas of the same application domain. For a very large number of such schemas, it is thus important to first categorize schemas by domain. Several approaches have been proposed for the automatic domain categorization problem of web forms (Barbosa 2007, Mahmoud and Aboulnaga 2010) typically based on a clustering of attribute names and the use of further features such as explaining text in the web page where the form is placed. While some approaches (Barbosa 2007) considered the domain categorization for only few predefined domains, (Mahmoud and Aboulnaga 2010) cluster schemas into a previously unknown number of domain-like groups that may overlap. This approach has also been used in (Eberius et al 2013) for a semantic grouping of related web tables. The categorization of web tables should be based on semantically enriched attribute information, the attribute instances and possibly information from the table context in the web pages (Balakrishnan et al 2015). The latter study observed that a large portion of web tables includes a so-called “subject” attribute (often the first attribute) indicating the kinds of entities represented in a table (companies, cities etc.) which allows already a rough categorization of the tables.

Directions for Future Research

Compared to the huge amount of previous work on pairwise schema matching, research on holistic schema matching is still at an early stage. As a result, more research is needed in improving the current approaches for a largely automatic, effective and efficient integration of many schemas and ontologies. In addition to approaches for determining an initial integrated schema/ontology, there is a need for incremental schema integration approaches that can add new schemas without having to completely re-integrate all source schemas. The use of very large sets of web-extracted datasets such as web tables also needs more research to categorize and semantically integrate these datasets.

To deal with a very large number of input schemas it is important to avoid the quadratic complexity of a pairwise mapping between all input schemas. One approach, which is especially suitable for integrating complex schemas and ontologies, is to incrementally match and integrate the input schemas, More research is here needed to evaluate how known approaches for pairwise matching and integration could be used or extended for such a setting.

For simpler schemas, or if only a clustering of concepts/attributes is aimed at, grouping of the input schemas by domain or mutual similarity is able to reduce the overall complexity since matching can then be restricted to the schemas of the same group. To limit the effort for matching schema elements within such groups performance techniques from entity resolution such as blocking (Koepcke and Rahm 2010) can further be utilized. More research is thus needed to investigate these ideas for a very large number of schemas / datasets.

References

- Balakrishnan S, Halevy AY, Harb B, Lee H, Madhavan J, Rostamizadeh A, Shen W, Wilder K, Wu F, Yu C (2015) Applying web tables in practice. Proc. CIDR
- Batini C, Lenzerini M, Navathe SB (1986) A comparative analysis of methodologies for database schema integration. ACM Comput. Surv. 18(4), 323–364
- Barbosa L, Freire J, Silva A (2007) Organizing hidden-web databases by clustering visible web documents. Proc. ICDE conf., 326–335
- Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32(suppl 1), D267–D270
- Das Sarma A, Dong X, Halevy AY (2008) Bootstrapping pay-as-you-go data integration systems. Proc. ACM SIGMOD conf.
- Do HH, Rahm E (2002) COMA – A System for Flexible Combination of Schema Matching Approaches. Proc. VLDB conf., 610–621
- Eberius J, Damme P, Braunschweig K, Thiele M, Lehner W. (2013) Publish-time data integration for open data platforms. Proc. ACM Workshop on Open Data
- Euzenat J, Shvaiko P (2013) Ontology Matching. 2nd edition, Springer
- Gross A, Hartung M, Kirsten T and Rahm E (2011) Mapping Composition for Matching Large Life Science Ontologies. Proc. ICBO, 109-116
- Gruetze T, Boehm C., Naumann F. (2012) Holistic and scalable ontology alignment for linked open data. Proc. LDOW
- He B, Chang, KC (2003) Statistical schema matching across web query interfaces. Proc. ACM SIGMOD conf., 217–228
- He H, Meng W, Yu CT, Wu Z (2004) Automatic integration of Web search interfaces with WISE-Integrator. VLDB J. 13(3): 256-273
- Hu W, Chen J, Zhang H, Qu Y (2011) How matchable are four thousand ontologies on the semantic web. Proc. ESWC, Springer LNCS 6643
- Koepcke H, Rahm E (2010) Frameworks for entity matching: A comparison. Data Knowl. Eng. 69(2), 197–210
- Limaye G, Sarawagi S, Chakrabarti S (2010) Annotating and searching web tables using entities, types and relationships. PVLDB 3(1–2), 1338–1347
- Madhavan J, Bernstein PA, Doan A, Halevy AY (2005) Corpus-based Schema Matching. Proc. IEEE ICDE conf.
- Mahmoud HA, Aboulnaga A (2010) Schema clustering and retrieval for multi-domain pay-as-you-go data integration systems. Proc. ACM SIGMOD conf.
- Nentwig M, Soru T, Ngonga Ngomo A., Rahm E (2014) LinkLion: A Link Repository for the Web of Data. Proc. ESWC (Satellite Events), Springer LNCS 8798
- Rahm E (2011) Towards Large-Scale Schema and Ontology Matching. In: Schema Matching and Mapping, Springer
- Rahm E (2016) The case for holistic data integration. Proc. ADBIS conf., Springer LNCS 9809
- Rahm E, Bernstein PA (2001) A Survey of Approaches to Automatic Schema Matching. VLDB Journal 10(4):334–350
- Raunich S, Rahm E (2014) Target-driven merging of taxonomies with ATOM. Information Systems, 42:1–14
- Saleem K, Bellahsene Z, Hunt E (2008) PORSCHE: Performance Oriented SCHEma mediation. Inf. Syst. 33(7-8): 637-657
- Wang J, Wang H, Wang Z, Zhu KQ (2012) Understanding tables on the web. Proc. ER conf., Springer LNCS 7532, 141–155

- Yakout M, Ganjam K, Chakrabarti K, Chaudhuri S (2012) Infogather: entity augmentation and attribute discovery by holistic matching with web tables. In: Proc. ACM SIGMOD conf., 97–108

Cross-References:
Large-scale schema matching,
Large-scale entity resolution