A SURVEY OF APPLICATIONS AND RESEARCHES ON SCHEMA MATCHING BETWEEN GIS SPATIAL DATA

WANG Yu-hong[a], ZHANG He-bing[a], XU Jun[b]

[a] School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China-wangyh@hpu.edu.cn
[b] School of Economics and Management, Henan Polytechnic University, Jiaozuo, China-xujun@hpu.edu.cn

KEY WORDS: Implementation Approach; Efficiency Optimization; Result Representation; Capability Evaluation

ABSTRACT:

As a fundamental problem of data management and application technology, schema matching has aroused the universal concern of the academic circles worldwide in recent years. In order to deepen the understandings of schema matching between spatial data and to identify its uses, the documentation method is adopted in this paper to firstly summarize and describe the foundation position and guidance role of schema matching in some typical applications such as spatial data integration (including schema-level integration and instance-level integration), updating information propagation, semantic query and handling, web geo-service finding. Then, aiming to the manual performance limitations of schema matching task in most systems, the previous works on schema matching are discussed mainly from four aspects of matching implementation approaches, matching efficiency optimization, matching results representation and matching capability evaluation for designing an automated approach and system. The related theories, models, approaches, limitations and new trends of current researches on schema matching are respectively analyzed. The conclusion is drawn by these analyses that schema matching researches are still faced with many theoretical and technological problems, the matching between schemas of spatial data will be more difficult and severe, and thus needs further studies since they are more heterogeneous, vaster and complex in structure than schemas of common data.

1. INTRODUCTION

Along with the increasingly maturation and widely popularization of GIS science and technology, GIS spatial data is rapidly increasing day by day. In order to take full advantage of these obtained data, to reduce the cost of system development and to promote their comprehensive analysis and application, the sharing and interoperation issues of spatial data are always the core and focus in the field of GIS study. The theoretical and technological problems associated with spatial data sharing and interoperation are very many, such as data schema integration (or merging), data instance integration, updating information propagation, semantic query processing, geo web service finding, and so on. Although these problems vary in the concrete solutions, there is a common key link during them, which is schema matching.

Schema Matching is the process of finding the semantical same or related elements from two or several data schemas based on various kinds of auxiliary decision-making information, and specifying the actual mapping relationships among them according to the application requirements. For example, the different levels of the related elements and their mapping relationships shown in the right part of Figure 1 can be found and specified by schema matching from partial schemas of two GIS databases shown in the left part of Figure 1.

In order to deepen the understanding of schema matching issue between spatial data and provide theoretical basis and technical reference for developing the efficient and practical schema matching systems, the typical applications of schema matching are firstly summarized in this paper, and then the related contents, principles, models and approaches achieved by the current researches are discussed.

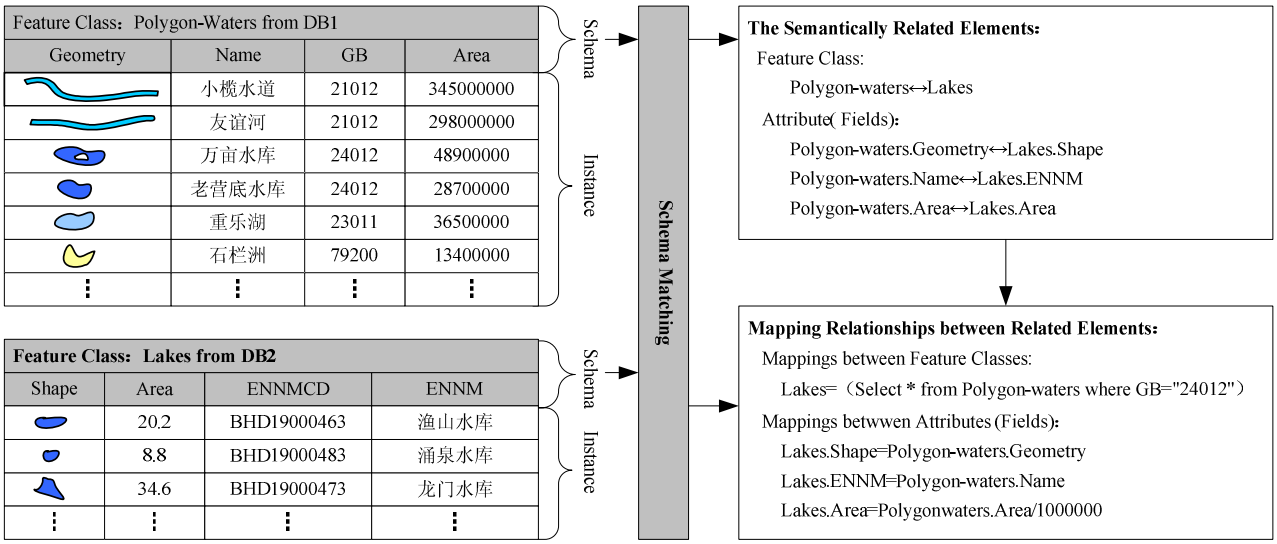

Figure 1. Diagram of Schema Matching

2. MAIN APPLICATIONS OF SCHEMA MATCHING

To motivate the importance of schema matching, we give the brief summaries of the applications of schema matching in GIS domains in the following section.

2.1 Spatial Data Schema Integration

Most work on schema match has been motivated by schema integration. Schema integration is the process of constructing of global schema from a group of independently developed schemas. Due to the difference in application fields, development habits and preferences, the schemas to be integrated may be different in logical structures and expression forms, even if they are used to describe the same phenomena or things. Thus, the first step of schema integration is to identify the related elements and specify the mappings between them through schema matching. Only according to these mappings some operations such as merging, eliminating redundancy and reorganizing can be performed on local schemas to produce a global comprehensive schema (Wang, et al., 2007; Volz, et al., 2008).

2.2 Spatial Data Instance Integration

Data instance integration is to organically combine the actual data records from various sources into a whole for transparently and seamlessly accessing and utilizing them. The core of data instance integration is to build schema mapping relationships by schema matching (Liu, et al., 2006). The required instances in data sources are filtered, extracted, transformed, fused, cleaned in term of schema mappings and uploaded into the target database or uniform retrieval interfaces for shielding the instance expression differences among data sources (Li, et al., 2012).

2.3 Updating Information Propagation

Updating information propagation is the process of utilizing the updating and change information of spatial entities (or features) in the newly-updated spatial database to revise, refine and correct the content of other databases constructed based on the original copy of it for ensuring that they also have an up-to-date representation of the real word. One of its basic requirements is to keep the updated databases autonomous, complete, correct and consistent as much as before. To meet this requirement, various operations, such as schema matching, change detection, entity identification, updates integration, etc, are proposed by many researchers (Laurent, 1998; Arnaud, et al., 2004; Wang, et al., 2010). The most important of these operations is schema matching because it is the basis for other operations and the results achieved by it can be used to guide and facilitate other operations.

2.4 Semantic Query Processing

Majority of spatial data query are based on keyword matching at present. If the keywords input by users are not all identical to the names of schema elements of the queried data, the not-needed or useless results will be returned. In order to overcome this defect, semantic query theory is proposed (Wang, et al., 2007). Semantic query, also called semantic retrieval, concept matching, refers to transform the keywords in query statements to make them uniform with the schema elements through matching operations between query statements and schema elements for returning the accurate results.

2.5 Geo Web Services Finding

Geo web services are some internet applications which can supply some basic geographical operations such as address matching, map drawing, route planning to developers and allow them to integration the spatial data and the related functions into their own web applications without implementing them by themselves (Bernard, et al., 2003). Along with the occurrence of more and more geo web services, it is particularly important to rapidly and accurately find the needed ones. There will no or unsatisfactory services to be found during services finding, once the service requestor and provider used the different terms to describe the same concepts or used the same term to describe the different concepts. Moreover, the semantic heterogeneities resulted from the version differences of geo web services will more increase the difficulty of service finding. Like semantic query, schema matching can be used to facilitate the solution of this question (He, et al., 2011).

3. EXISTING RELATED RESEARCHES ABOUT SCHEMA MATCHING

As a fundamental problem of data management and application technology, schema matching has aroused the universal concern of the academic circles worldwide in recent years. A lot of research works on it have been carried out by people from various fields such as database, artificial intelligence, information retrieval, knowledge management, semantic web and so on. To sum up, the existing researches mainly focus on four aspects: implementation approach, efficiency optimization, result representation, capability evaluation.

3.1 Implementation Approach

Currently, schema matching is typically performed manually, perhaps supported by a visual interface such as attribute transfer mapping functions in ArcGIS 10.0, workbench component in FME2011, etc. The manual specifying of schema matches is assumed that users have sufficient knowledge of both source and target schema. Moreover, it tends to a tedious, time-consuming, error-prone, and therefore expensive process along with the increase of the number of elements to be matched.

To overcome the limitations of manual matching, various automated (or semi-automated) approaches are proposed. The basic idea of the automated approaches is to evaluate and express the similarity of elements between schemas. If a certain degree of similarity can be detected, two elements can be assigned to each other. According to the source of information that can be used for similarity evaluation, the automated approaches can be divided into two main kinds: element-based approach and instance-based approach.

Element-based approach determines schema matches by comparing information on elements themselves (such as names, documents, specifications) based on prerequisite that similar elements may have similar representations. String-matching methods and linguistic tools are used to measure the similarities among class names and attribute names (Stephen, et al., 1990). Descriptions of classes and attributes in design documents can be compared through document-similarity measures developed in the information-retrieval field (Benkley, et al., 1995). Specifications (including data type, length, value range, optional, etc) on attributes can be compared according to the predefined rules (Li, et al., 1993; Qiang, et al., 2003).

However, element-based approach is often prone to produce the imperfect results due to the unreliability and incompleteness of element-level information. For example, two elements that share the same name can refer to different classes or attributes; two elements with different names can refer to the same class or attribute. There may be several elements with similar specificati -ons, but different meaning between two schemas. Design documents are often outdated, incomplete, incorrect, ambiguous, or simply not available.

Instances can give important hints about the contents and meaning of schema elements (Erhard, et al., 2001), and thus they are usually used for attribute-level matches based on the simple principle: similar statistical characteristics or data values between attributes imply corresponding attributes. For example, summary instance information of attributes (such as mean and standard deviation, max, min, average, etc) is used together with the specifications to measure attribute similarities (Li, et al., 1993; Qiang, et al., 2003). However, summary instance information is necessary but not sufficient for schema matching (Cecil, et al., 2003). In (Lu, et al., 1997), the statistical correlation coefficients between numerical attributes is firstly computed based on the overlapping instances and then is used as attribute similarity measures. In (Cecil, et al., 2003; Bilke, et al., 2005), the literal similarity among the overlapping instances is used to measure similarity between textual attributes.

However, because of the difficulties of comparing and analyzing data instances with the unknown schema matches and the diversities of instance representation (Zhao, 2007), the instance-based approaches are still facing with at least three problems. Firstly, they only focus on attribute-level matches and leave the problem of determining class-level matches unsolved. Secondly, the overlapping instances are often obtained manually or by comparing common ID between objects. This requires that common ID must exit and has been matched. However, common ID usually does not exist. Thirdly, some real attribute-level matches will be missed due to the lake of the comprehensive consideration to the discrepancies between the overlapping instances, such as different formats, different scales, spelling errors, different code, etc.

3.2 Efficiency Optimization

The current difficulty of schema matching lies not only in the lack of practical strategies or rules to identify whether the schema elements are matched but also in the high cost for performing matching based on the predefined rules, which generally need a large number of computations and comparisons to find the possible matches. Therefore, the researches on schema matching efficiency optimization models and algorithms have to be strengthened. There are only a few systems considering or handling the problem of performance efficiency. To sum up, the following four techniques for improving the performance were introduced in the different types of systems (Eric, et al., 2010).

Divide and conquer: A number of systems apply a divide and conquer strategy when matching large schemas. They first try to manually or automatically identify relevant fragments, blocks, partitions or clusters. The further matching is then performed on these identified schema parts, which reduces the search space. Unfortunately, this approach could worsen the overall result quality.

Filtering schema parts: Some systems apply a schema reduction upfront by filtering out the relevant context or by involving the user through a questionnaire. Some systems automatically identify non-needed edges in the schema-graph structure or apply heuristics to reduce the number of comparisons at the cost of quality. Also the famous edit-distance algorithm can be improved by early pruning of comparisons. Similar strategies for reducing the search space were proposed in the record linkage area. These strategies are called blocking and try to reduce the number of candidate record comparison pairs while still maintaining reasonable linkage accuracy.

Avoiding repetitions: A general performance technique is to avoid the repeated execution of the same subtask. For example, a pre-matching step such as tokenizing all labels avoids the repeated tokenization in later match comparisons.

Improved data structures: A number of techniques use special data structures like indexes or hash tables to improve performance. Indexing helps to quickly identify the right elements to compare with. For instance, the B-Match-Approach indexes tokens and its labels. That saves string comparisons based on the assumption that two similar labels share at least one common token. Others remove the nested looping effort since each element in the source needs to be compared to each element in the target by introducing a hash-join like method. They also cache already computed results for later reuse.

3.3 Result Representation

The Main task of matching results representation is to organize and store the related schema elements and the actual mapping between them achieved by schema matching, and to build the necessary access and retrieval approaches for guiding and simplifying other operations in various applications. At present, some matching tools directly store results into plain text files according to their own needs. This kind of representation lacks the sufficient semantical expressiveness and processing capability, and thus makes it complicated to access matching results and difficult to sharing them among many systems. Several tools store matching results into relational database. Due to the semi-structured characteristic of matching results, the relational expression will yield many redundant fields with “null” values in result tables and be incapable of effectively recording some complex matching or mapping relationships such as conditional matches, partial matches and computational matches (Han, et al., 2006). Moreover, as the schema elements to be matched change, the structure of result tables may also change accordingly. This makes it inconvenient to manage and maintain the matching results.

Aiming to limitations of the above representation ways, some researchers have tried to utilize logic-based languages or semi-structured models to represent and store matching results. For example, Yuan et al. (2005) used first order logic to express the complex mappings between XML schemas and OWL ontologies. In order to provide a better understanding of the commonalities and differences of existing proposals for ontology mapping languages, Serafini et al. (2005) used distributed first order logic (DFOL) to give a formal comparison of existing mapping languages. In BRICK system (Kearney, et al., 2007), XML model is used to store and manage ontology mappings.

3.4 Capability Evaluation

To identify a solution for a particular match problem, it is important to understand which of the proposed techniques performs best. The performance capabilities of matching systems incorporate different (possibly conflicting) aspects such as effectiveness, efficiency, genericity and usability. The effectiveness is concerned with the accuracy and the correctness of the matching results. The efficiency is concerned with resources consumption (time, memory,...). The genericity is concerned with the application domains of matching systems. The ideal matching methods should be applicable to different match tasks from various domains and for different data models. The usability is concerned with the ease-of-use of match systems and the manual effort savings (Do, et al., 2002; Algergawy, et al., 2008; Köpcke, et al. 2010).

At present, the evaluation of matching capabilities mainly concentrates on the effectiveness aspect. To show the effectiveness, some researchers have usually demonstrated their proposed tools to some real world scenarios or conducted a study using a range of schema matching tasks. However, it is quite difficult to evaluate the matching systems for several reasons. First, the systems are not always available as a demo and it is not possible to test them against specific sets of schemas. Second, some systems require specific resources to be efficient, like an ontology or a thesaurus, which are not always available. Finally, some matching tools take as input specific additional files (Duchateau, et al., 2007a). Thus, a generally accepted benchmark is particularly important to users, developers and researchers for comprehensively comparing and evaluating them (Zohra et al., 2011). Some useful attempts and prototypes (e.g. XBenchMatch, STBenchmark) have been done and developed (Duchateau, et al., 2007b; Bogdan, et al., 2008).

4. CONCLUSIONS

After more than 30 years of unremitting efforts, gratifying results were achieved in the studies of schema matching, from the simple matcher based on information from schemas themselves, to the composite matcher utilizing various information such instances, structures, to the human-based matching tools or systems and to the systematic theory supports Schema matching still is a challenging issue due to its subjectivity, uncertainty and complexity.

According to the documents and materials, the researches on schema matching in the field of GIS spatial data are currently very weak. The existing relevant discussions are mostly parenthetic explanations on schema matching concept and lack the pertinent and detailed analysis despite of a few works concentrating on the design and realization of the actual approaches and systems. Compared with the characters of spatial data schema of many types, large scales and complex structures, the current researches are not sufficient enough to meet the requirements of an ideal system on genericity, robustness, flexibility, interactivity and so on. It is very necessary to actively and deeply carry our the further research works on schema matching so as to provide theoretical supports and technical guarantees for effective sharing and intelligent services of spatial data resources.

ACKNOWLEDGEMENTS

The work described in this paper is supported by the united fund project of Natural Science Foundation of China and the People's Government of Henan Province (U1304401) and the sub-project of the 12th Five Years Programs for Science and Technology Development of China (2012BAJ23B04-2).

REFERENCES

Algergawy A., Schallehn E., Saake G., 2008. Combining Effectiveness and Efficiency for Schema Matching Evaluation. *Proceedings of 1st International Workshop on Model-Based Software and Data Integration*, Germany, pp.19-30.

Arnaud Braun, 2004. From the Schema Matching to the Integration of Updating Information into User Geographic Database, *Proceeding of Geoinformaticas* 2004, pp.211-218.

Benkley, Fandozzi, Housman, Woodhouse, 1995. Data element tool-based analysis (DELTA). *Technical Report MTR 95B0000147*, The MITRE Corporation, Bedford.

Bernard L., Einspanier U., Lutz M., et al., 2003. Interoperability in GI Service Chains — The Way Forward. *The 6th AGILE Conference on Geographic Information Science*, Lyon , France.

Bilke Alexander, Naumann Felix, 2005. Schema Matching using Duplicates. *Proceedings of the 21st International Conference on Data Engineering*, pp.69-80.

Bogdan A., Tan W., Velegrakis Y., 2008. STBenchmark: Towards a Benchmark for Mapping Systems. *Proceeding of VLDB '08*, August 23-28, Auckland, New Zealand, pp.230-244.

Cecil E. H. C., Roger H. L. C., Ee-Peng L., 2003. Instance-based attribute identification in database integration, *VLDB Journal*, Vol.12, No.3, pp.228–243.

Do H., Melnik S., Rahm E., 2002. Comparison of schema matching evaluations. *Lecture Notes in Computer Science: Web, Web-Services, and Database Systems*, 2593, pp.221-237.

Duchateau F., Bellahsène Z., 2007b. Designing a Benchmark for the Assessment of XML Schema Matching Tools, *Proceeding of VLDB 2007*, September 23-28, Vienna, Austria.

Duchateau F., Bellahsène Z., Hunt E., 2007a. XBenchMatch: a Benchmark for XML Schema Matching Tools. *Proceeding of VLDB 2007*, September 23-28, Vienna, Austria, pp.1318-1321.

Erhard Rahm, Philip Bernstein, 2001. A survey of approaches to automatic schema matching. *VLDB Journal*, Vol.10, No.4, pp.334-350.

Eric P., Henrike B., Erhard R., 2010. Rewrite Techniques for Performance Optimization of Schema Matching Processes. *Proceeding of 13th International Conference on Extending Database Technology*, Lausanne, Switzerland, pp.433-464.

HAN Zhongming, CHEN Dehua, LE Jiajin, 2006. Schema Mapping and Representation. *Journal of Huadong University*, 22(2), pp.42-45.

HE Jie, CHEN Neng-cheng, WANG Wei, et al., 2011. A uniform approach for multi-version web feature service retrieve based on dynamic schema matching. *Science of Surveying and Mapping*, Vol. 36, No.1, pp. 169-172.

Kearney K., 2007. Ontology Mapping in BRICKS. *Proceedings of Workshop on Ontology-Driven Interoperability for Cultural Heritage Objects*.

Köpcke H., Rahm E., 2010. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2), pp.197-120.

Laurent Spery, 1998. Spatial Data Transfer in the case of Update. *International Archives of Photogrammetry and Remote Sensing*, vol.32, no.4, pp.586-593.

LI Jun, SU Guo-zhong, LI Meng, 2012. Shielding the heterogeneity of geospatial data sources by using GML schema mapping. *Science of Surveying and Mapping*, Vol. 37, No.1, pp. 38-41.

Li Wen-Syan, Clifton Chris, 1993. Using field specifications to determine attribute equivalence in heterogeneous databases. *Proceeding of Third International Workshop on Research Issues in Data Engineering - Interoperability in Multidatabase System*s, Vienna, Austria, pp.174-177.

LIU Min-chao, LIU Wei-dong, 2006. Research on key problems in data integration system. Journal of Computer Applications, Vol.26, No.7, pp.1507-1510.

Lu Hongjun, Fan Weiguo, Cheng Hian Goh, 1997. Discovering and Reconciling Semantic Conflicts: A Data Mining Perspective. *Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics*, Leysin, Switzerland.

Qiang Baohua, Wu Kaiwei, Wu Zhongfu, 2003. A data-type-based approach for identifying corresponding attributes in heterogeneous databases, *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, Xi'an, pp.299-344.

Serafini L., Stuckenschmidt H., Wache H., 2005. A formal investigation of mapping languages for terminological knowledge. *Proceedings of the 19th international joint conference on Artificial intelligence*, pp.576–581.

Stephen Hayne, Sudha Ram, 1990. Multi-user view integration system (MUVIS)： an expert system for view integration, *Proceedings of the Sixth International Conference on Data Engineering*, pp.402-409.

Volz S, Danielas N, Grossmann M, et al, 2008. On Creating a Spatial Integration Schema for Global, Context-aware Applications. *Proceedings of GeoInfo 2008*, pp.13-24.

WANG Hongding, TAN Shaohua, TANG Shiwei, etc., 2007. Schema Merging Study with Semantic Relationships of Schema Elements. *Acta Scientiarum Naturalium Universitatis Pekinensis*, Vol.43, No.3, pp405-411.

WANG Yandong, GONG Jianya, DAI Jingjing, 2007. Spatial Data Semantic Query Based on Ontology. *Journal of Geomatics*, Vol.32, No.2, pp.32-35.

WANG Yu-hong, CHEN Jun, 2010. Implementation Approach for Propagating Updates of Fundamental Geographic Database. *Geomatics and Information Science of Wuhan University*, Vol. 39, No.1, pp.1116-1120.

Yuan A., Borgida A., Mylopoulos J., 2005. Constructing Complex Semantic Mappings between XML Data and Ontologies. *Proceedings of the 4th International Semantic Web Confrence*, Ireland, pp.6-19.

Zhao Huimin, 2007. Semantic matching across heterogeneous data sources. *Communications of the ACM*, Vol.50, No.1, pp.45-50.

Zohra B., Angela B., Fabien D., et al., 2011. On Evaluating Schema Matching and Mapping. *Schema Matching and Mapping*, Springer Berlin Heidelberg, pp.253-291.