8-6-2011

A Survey of Cognitive Theories to Support Data Integration

Roman Lukyanenko
Memorial University of Newfoundland, roman.lukyanenko@mun.ca

Joerg Evermann
Memorial University of Newfoundland, jevermann@mun.ca

A Survey of Cognitive Theories to Support Data Integration

Roman Lukyanenko
Memorial University of Newfoundland
roman.lukyanenko@mun.ca

Joerg Evermann
Memorial University of Newfoundland
jevermann@mun.ca

ABSTRACT

Business intelligence applications are being increasingly used to facilitate managerial insight and maintain competitiveness. These applications rely on the availability of integrated data from multiple data sources, making database integration an increasingly important task. A central step in the process of data integration is schema matching, the identification of similar elements in the two databases. While a number of approaches have been proposed, the majority of schema matching techniques are based on ad-hoc heuristics, instead of an established theoretical foundation. The absence of a theoretical foundation makes it difficult to explain and improve schema matching process. This research surveys current cognitive theories of similarity and demonstrates their application to the problem of schema matching. Better integration techniques will benefit business intelligence applications and can thereby contribute to business value.

KEYWORDS

Schema matching, human information processing, business intelligence, database integration, relational database, similarity, cognitive psychology

INTRODUCTION

With the proliferation of information systems in business, and explosion of the data sources used to support modern business operations, the need to have a consistent view of data is ever increasing. In particular, business intelligence (BI) draws on integrated data from multiple transactional data sources to provide business insight through analytics and data mining techniques. Since much of the corporate information is stored in multiple relational databases, their integration is critical if organizations are to derive business value from business intelligence applications. Database integration may also arise as the result of a business merger, internal organizational restructuring or external partner integration. Database integration can also be used as a way to synchronize enterprise applications, such as website, data warehouse and an accounting system for one company, powering decision support and management BI applications. Hence, the quality of the database integration affects business processes and the overall competitiveness of a company.

In the context of data integration for BI, as well as in other contexts, the central objective of data integration is matching those elements of a database that are similar or identical. This is called *schema matching* and is the focus of this paper. Despite the importance of database integration, it remains an "extremely difficult problem" (for review of the discipline see Doan and Halevy, 2005; Rahm and Bernstein, 2001). Without access to the database designer, it is often difficult to understand the meaning of a given database element, especially considering the widespread use of acronyms, idiosyncratic abbreviations (Evermann, 2008b) and idiosyncratic conceptual models of the same domain (Parsons and Wand, 2000). We note that while in the wider process of data integration, there are BI specific issues to consider, e.g. the data warehouse structure or the ETL (extraction-transformation-load) process, the core step of schema matching is described in the literature as context independent (Doan et al., 2005; Rahm et al., 2001), though a recent study by Evermann (2010) appears to challenge this.

Several studies (Evermann, 2008a, b) referred to similarity judgment as a promising theoretical foundation for schema matching, yet there has been no clear and consistent account of how similarity theory can be applied to the challenges of database integration. This paper proposes similarity theory as a promising theoretical foundation for database integration. The main contribution of this paper is a review of current theories of similarity and a proposal of how they might be applied to schema matching.

The remainder of the paper is organized as follows. The next section briefly describes the process of database integration with a focus on schema matching. This is followed by a brief historical review of theories of similarity. The main section then examines current theories of similarity and shows how each can be applied to schema matching. The paper concludes with a general discussion and outlook to future research.

A Brief Review of Database Integration

Database integration is a multistep process of finding similar entities in two or more databases to create a non-redundant, unified view of all databases (Batini, Lenzerini and Navathe, 1986). A number of approaches to database integration exist in

two categories: methods that use schema-level information and those that use instance information (Evermann, 2008a; Evermann, 2009).

Ideally, any relevant information (the database designer's knowledge, documentation, application software and the database schema itself) should be used to determine if two elements are similar. Not all of that information may be accessible, and typical schema matching research has narrowed the scope to: (1) schema and entities that are described by it, such as tables, views, fields, relationships (Agarwal and Karahanna, 2000; Bin and Che-Chuan Chang, 2006; Chen, Promparmote and Maire, 2006; Domshlak, Gal and Roitman, 2007; Evermann, 2009), (2) the data records themselves (Fan, Lu, Madnick and Cheung, 2001; Kang and Naughton, 2003; Miller, Hernandez, Haas, Yan, Ho, Fagin and Popa, 2001) (3) and the immediate database environment: applications that connect to the database including the interface, programming code and embedded business logic (Evermann, 2010). While the prevailing view considers schema central to database integration, several studies combined schema with instance or environmental information for a holistic approach to integration (Evermann, 2010; Zhao and Ram, 2007).

Both the theory and practice of database integration has used a variety of approaches, ranging from adopting the theories of meaning (Evermann, 2008b) to employing heuristics without extensive theoretical support (Berlin and Motro, 2001; Imhoff, 2005; Shen, Vemuri, Fan and Fox, 2008).

The outcome of the integration can be evaluated by the partially formalized criteria for successful schema matching that have been proposed by Batini et al. (1986). These include completeness and correctness (all elements of source schemas need to be represented in the unified schema), minimality (the integrated schema needs to represent the same concept once no matter how often it appears in source schemas), and understandability (the unified schema needs to be easy to understand both for the designer and for the application users). The research on database integration consistently acknowledged the challenges involved in integration, which offers a strong motivation for the development of the similarity-based theory.

To relate theories of similarity to specific challenges of schema matching we use a running example of two fictitious databases (see Figure 1). Both databases represent the same domain and have approximately the same scope. The databases were designed following empirical findings from (Evermann, 2010) where a number of database elements have been identified by integration experts as salient for successful schema matching. These elements include schema structure (tables and relationships), constraints (keys, types, field sizes), object names (field and table names) and the instances themselves (see: Evermann, 2010).

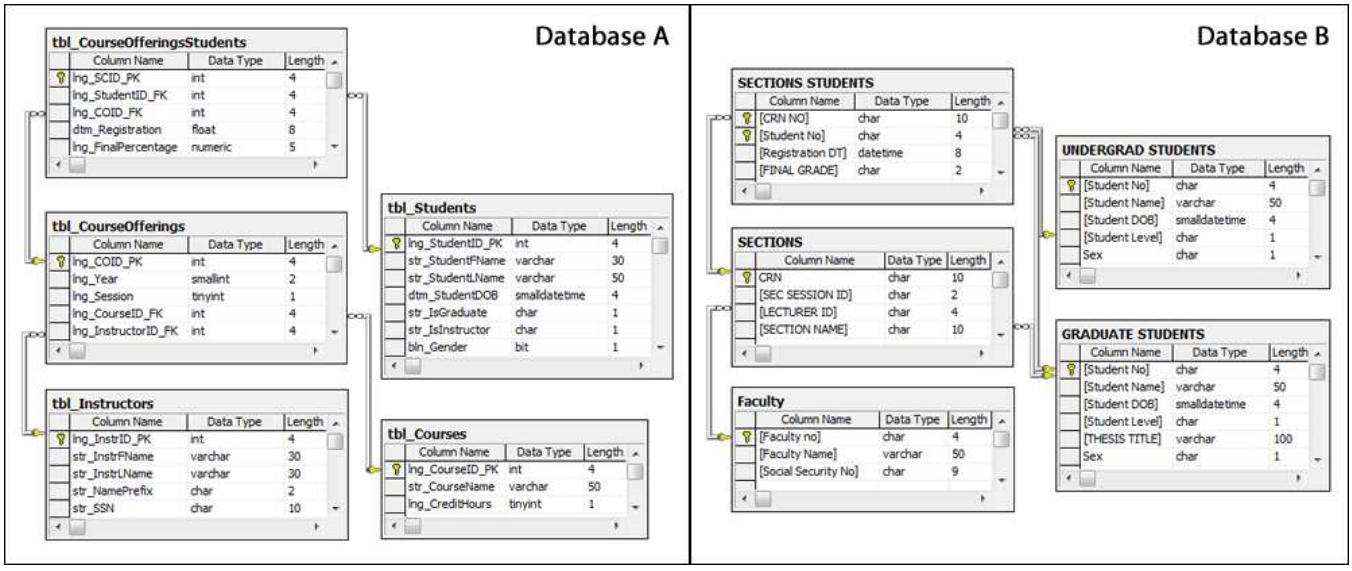

Figure 1.Two hypothetical databases belonging to the same domain that require integration

Theories of Similarity in Database Integration

Integrating any two sources requires passing a judgment about the degree of their similarity. Similarity, as one of the basic cognitive processes (Imai, 1977; Rumelhart and Abrahamson, 1973), has seen a wide range of applications from multimedia

querying and image databases (Santini and Jain, 1999) to conceptual modeling and schema integration (Evermann, 2009). As Yi et al. (2008) point out, "similarity may be the most universal relationship that exists between any two objects" (p. 743).

Early research (Sjöberg, 1969) viewed the possibility of measuring similarity with reservation: "[t]he basic problem is whether it is at all possible to measure the similarity (of pairs of objects) for one individual on one occasion" (p. 441). Measuring similarity remains a challenging task, and in the context of schema matching one study that did not detect universal patterns concluded: "people differ strongly in their similarity judgment" (Evermann, 2008a). Later research, discussed below, proposed a variety of measurements of similarity and assumed the possibility of a general theory for similarity judgment (Ashby and Perrin, 1988; Hahn, Chater and Richardson, 2003; Hahn, Close and Graf, 2009).

Over the last four decades, a multitude of similarity theories have been proposed in the field of cognitive psychology (Schwering, 2005). Similarity theories can be divided into three broad categories: spatial, featural (that has a powerful off-spring theory of structural mapping), and transformational.

Representing similarity as a spatial relationship between stimuli was one of the first influential approaches to similarity judgment (Rumelhart et al., 1973; Shepard, 1962). Similarity was seen as a "degree of proximity" (Shepard, 1962) of one object from another in a psychological space. The challenge of the theory was to convert an implicit distance between two concepts into an explicit one that could be modeled and computed. The final result was a set of coordinates in a multidimensional Euclidean space that represents a mental model of the relationship between two or more elements. The dimensions in spatial theories can represent different relationships and thus, distances between objects may vary, depending on which relationships are considered. Thus, while dimensions address the issue of variability of distance, the practical application and implementation of this theory remains unclear. A number of researchers question adequacy of spatial approach to similarity (Ashby et al., 1988; Hahn et al., 2003; Tversky, 1977).

This research focuses on those theories that are the most influential and have the most applicability to schema matching, thus we do not consider the spatial approach to similarity any further. While there are recent publications representing each of the three frameworks, transformational and featural theories have seen the most development, producing a multitude of variations.

TRANSFORMATIONAL THEORIES OF SIMILARITY

Using the stimuli of two sets of black and white beads, Imai (1977) proposed similarity as a sequence of transformations that are necessary to convert one object into another (Imai, 1977). The types of transformations included mirror-image, phase, reversal, and wave length. Imai concluded that the abstract nature of the beads showed that predictions of similarity are possible "without referring to the identity of the features" (Imai, 1977). Such semantic abstraction can be exploited by automatic schema matching techniques. Many database elements are named using abbreviations and cryptic acronyms; and extracting meaning may not be always possible (Evermann, 2008b).

Building on Imai's findings, Hahn et al (2003) developed the representational distortion (RD) transformational model, designed to overcome "fundamental limitation" of other major similarity theories. Natural phenomena are complex, and interactions between parts of complex objects are ignored by similarity theories that examine simple words, sentences, or shapes. Yet reasoning about reality requires "not just specifying what features it has, but, crucially, how they are interrelated" (Hahn et al., 2003). Similarity, therefore, is a function of the "complexity" required to "distort" or "transform" one object or concept into another. Formally, a distance function can be defined as the shortest algorithm necessary to transform A into B expressed as K (B | A), where K is a conditional Kolmogorov complexity. For entirely dissimilar objects, RD requires deletion of the original object and the reconstruction of the target *ab initio*, thus resulting in a high number of transformations.

Transformational theory applies to schema matching in multiple ways. First, it supports programming implementations, since transformational complexity can be viewed as the complexity of algorithms comprising data transformations, e.g. those typically used in such BI applications as ETL (extraction-transformation-load)(e.g. Simitsis and Vassiliadis, 2008). Second, as shown by (Hahn et al., 2003; Hahn et al., 2009; Hodgetts, Hahn and Chater, 2009; Imai, 1977), the internal composition of objects can be examined without necessarily considering their meaning. For example, when comparing two tables, the values of their attributes can be converted into conditional complexity. Modern ETL software provides complex scripting of data transformation, originally developed for the data warehousing context. Complexity-based similarity can be measured by the number of steps needed to merge two elements using such software (e.g. Pentaho, www.pentaho.com). For example, in database A, the column *bln_Gender* defines "male" as "0" and "female" as "1." Database B has two *char* fields "sex" with a size limit of one character and expected values of "M" and "F" specified by the constrains for each field. Mapping between 0 and "M" and 1 and "F" can be done using a string replacement transformation and then merged with stream from database A for the output (see Figure 2). Similarly, grades in database B (which are represented by letters A-F) can be translated using

ETL's range transformation to match the numeric scale defined in database A. Other parameters that define each column (default value, precision, scale, formulas) can also be compared by the steps necessary to match them with corresponding ones in another database. Comparing these measurements can offer guidance in automated decisions to treat certain columns of a database as similar and potentially integratable.

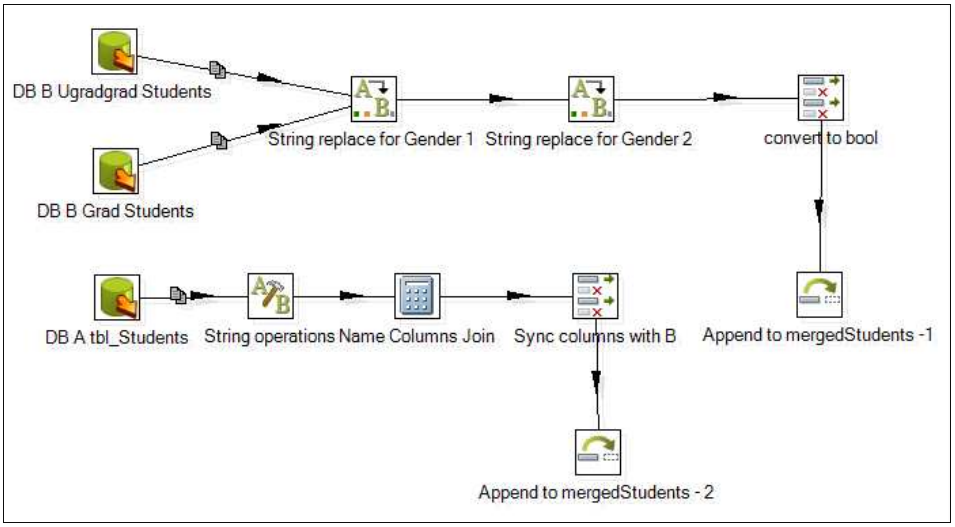

Figure 2: Flow diagram of transformations of three student tables from databases A and B into a common repository (Example uses the Pentaho Data Integration software, www.pentaho.com)

A recent addition to the RD model is a notion of transformation directionality, which states that object A may be more similar to object B than vice versa (Hahn et al., 2009). Transformations from one end point to another have different measures of complexity, the finding adding to a large body of research exposing asymmetrical nature of similarity (Gentner and Markman, 1997; Tversky, 1977). One source of asymmetry is the notion of typicality of an object (Rosch, 1975; Rosch, 1978). Categories are not uniform, and some members of categories are more typical than others. This asymmetry has an impact on schema matching. For example, this may mean that a table *Graduate Students* in database B is more similar to the table *tbl_Instructors* in database A (graduate students also teach courses) rather than the other way around. Asymmetric transformational complexity suggests it could be easier to transform a concept of a graduate student into a concept of an instructor than vice versa. The impact of directionality is seen when merging *str_InstrFName*, *str_InstrLName* and *str_NamePrefix* in database A with *Faculty Name* in database B. The decision to merge three columns from database A into a common *Instructor Name* requires significantly fewer ETL transformations than splitting *Faculty Name* in database B into three or more fields, especially considering complexity needed to capture prefixes, suffices, extended middle names and conjunctions. A complexity-based automatic algorithm approaching two elements from the wrong direction, therefore, may misrepresent their relationship and potentially trigger an incorrect merging decision.

STRUCTURAL THEORIES OF SIMILARITY

Possibly the most influential idea that built the foundation of modern research was the featural-contrast model developed by Tversky (1977). Tversky argued that the similarity of two objects, *a* and *b*, is a function of the number of features they have in common, and the number of features that are different in each object: $s(a,b) = f(A \cap B, A - B, B - A)$. Tversky also advanced the idea of *asymmetry* in similarity judgment. Following research on basic categories (Rosch, 1975) and prototyping (Rosch, 1978), Tversky reasoned that some objects are more prototypical than others. Thus, ovals will be similar to circles, but circles will be less similar to ovals. Tversky was also one of the first to explore the role of the frame of reference for feature matching (Tversky, 1977). He reasoned, for example, an essay could be similar to a fish if we know we are comparing their structures (head, body, and tail). The concept of a frame of reference is further developed by Christie and Gentner as part of the systematicity theory (Christie and Gentner, 2010).

Tversky's feature-contrast theory became the foundation of many structural approaches to similarity (Markman and Gentner, 1993). Accepting the featural model of similarity, Gentner (1983) saw a limitation of Tversky's theory in its reliance on attributes, and suggested that relations between attributes are often more salient in similarity judgment. Focusing on relations

and structure (Gentner et al., 1997; Markman et al., 1993), further proposed two fundamental assumptions of structural similarity: one-to-one matching and parallel connectivity.

One-to-one matching reflects a cognitive predisposition to map one element of the source to at most one element of the target. One-to-one matching is an important claim supported by a large body of cognitive research (Gentner et al., 1997; Holyoak and Thagard, 1989; Markman et al., 1993; Spellman and Holyoak, 1992; Yi, Xiaolong and Qiang, 2008). The assumption of one-to-one mapping can significantly reduce the complexity of schema matching problem.

It is not always possible, however, to find one-to-one correspondence between all elements. For example, three fields (*str_InstrFName*, *str_InstrLName* and *str_NamePrefix*) in the database A map to one field (*Faculty Name*) in the database B. This one-to-many mapping will tend to decrease the similarity between the table *tbl_Instructors* and *Faculty* (Goldstone and Medin, 1994).When a one-to-one mapping between features cannot be established, structural relations may guide similarity judgment (Spellman et al., 1992). The table *tbl_Instructors* exists in the same relationship with the remaining tables in database A (via table *tbl_CourseOfferings)* as does the table *Faculty* (via *Sections*) in database B. This similarity is based on the similarity of the overarching relation (teaching). Such alignment of elements based on the globally consistent structure reflects the concept of *parallel connectivity*. Parallel connectivity emphasizes similar relational structures between corresponding objects and features.

The natural human tendency to seek global consistency and coherence ensures that only the relationships that belong to a system of interconnected knowledge are used in mapping, while others are ignored. Gentner and Toupin (1986)write: "adults focus on shared systematic relational structure in interpreting analogy. They tend to include relations and omit attributes in their interpretations of analogy, and they judge analogies as more apt if they share systematic relational structure"(p. 282).The idea that people seek coherent knowledge and prefer it over independent information is the *principle of systematicity, which is well supported* in cognitive research (Yi et al., 2008).

STRUCTURE MAPPING ENGINE

Like the transformational approach, the structural account of similarity can be easily operationalized and applied to schema matching. One implementation is the Structure Mapping Engine (SME) (Falkenhainer, Forbus and Gentner, 1989; Lovett, Gentner, Forbus and Sagi, 2009a). SME is an iterative algorithm designed to find structurally relevant attribute matches between source and target. SME starts by identifying congruent predicates and roles in two domains. Predicates include relations, roles and attributes. Some of these matches may be inconsistent, and some accidental, not reflecting deeper meaning. For example, the field *str_CourseName* of *tbl_Courses* in database A appears similar to two fields *Student Name* in database B (both contain "name" and are of type *varchar* with size 50). A similarity score will be assigned to each match. Yet, clearly, this similarity does not mean that the table *tbl_Courses* should be matched with either or both of the *Student* tables in database B. This is ensured by the global alignment stage of SME. Global alignment examines existing local matches for an overall consistency and aims to maximize the *systematicity bias* (Lovett, Tomai, Forbus and Usher, 2009b). At this stage, many "superficial" matches may be rejected. Thus table *tbl_Instructors* should be matched with *Faculty* due to the similarity of roles and relations in the overall schema. After that, the corresponding elements of the two entities (term used in SME), will be reevaluated to reveal their deeper similarity (Markman et al., 1993). Following cognitive theory, we then expect *str_InstrFName*, *str_InstrLName* and *str_NamePrefix* in database A to be mapped to *Faculty Name* in "B." SME has been employed for a variety of tasks, such as conceptual properties of sketches (Forbus, Usher and Tomai, 2005), stories (Gentner, Rattermann and Forbus, 1993), geometric shapes (Lovett et al., 2009b). Schema matching may be another interesting application of the SME.

SIMILARITY AS INTERACTIVE ACTIVATION AND MAPPING (SIAM)

The procedural aspect of similarity judgment is explored in SIAM (Similarity as Interactive Activation and Mapping)(Goldstone et al., 1994). SIAM is based on a network of interconnected nodes, each of which represents the similarity between two mapped elements (features, objects or roles). Two or more nodes that represent high similarity of their mapped elements excite each other, while the nodes that represent low similarity between their mapped elements inhibit one another. The alignment process begins with the alignment of features to produce feature-to-feature nodes.

Although with a degree of discretion, SIAM contains constructs that can be mapped to basic elements of a database schema. Theoretical *features* can be operationalized as table attributes. Thus, the first step would be to map pairs of similar attributes. For example, the attribute *lng_StudentID_FK* can be mapped to *Student No*, and *dtm_Registration* to *Registration DT*. Incorrect matches are possible as well, since initial matches are based on surface similarity. As a database integrator spends more time examining the feature-to-feature mappings, discoveries of new similarities (such as similar data types) will place them in stronger correspondence. Following that, objects (database tables) possessing those features are examined for similarity and the focus shifts to object-to-object nodes. Thus, the table *tbl_CourseOfferingsStudents* will be matched with

Sections Students, and *tbl_CourseOfferings* with *Sections*. Discoveries of matches at the object-to-object nodes further excite nodes representing matches of features that belong to those objects. A sense of similarity grows stronger, as structural consistency begins to emerge. Any featural misalignments (one column in database B matching to several potential columns in database A) can now be reexamined and possibly resolved (Larkey and Markman, 2005). Structural alignment is further increased once role-to-role nodes are considered. Roles can be operationalized as foreign key relationships and reflect the position of tables in the overall schema. At this point, the table *tbl_Instructors* in database A can be placed in strong correspondence with the merged tables *Graduate Students* and *Faculty* in database B due to the overall similarity of their roles in the schemas. Role-to-role and the other node types mutually influence each other and either strengthen or weaken the similarity evaluations established during prior steps. Thus, based on the role-to-role alignment, fields from table *Graduate Students* will be placed in the correspondence with fields in the table *tbl_Instructors,* despite any initial (surface) dissimilarity between them.

During the alignment process, matches may occur between elements because of surface similarity (column *lng_StudentID_PK* in *tbl_Students* and *Student No* in *Student Sections*). Such matches are called "match out of place" (MOP) because they belong to dissimilar objects (tables). "Matches in place" (MIPs) are those that belong to similar objects and those objects share similar roles. The relative importance of MIPs compared with MOPs increases with processing time. According to Goldstone et al. (1994), over time, qualitative shifts in similarity occurs: "[e]arly in processing, local matches strongly influence similarity; with time, global consistency becomes more important" (p. 29). SIAM draws attention to the importance of the temporal aspect of similarity.

Discussion and Conclusion

In this paper, we have proposed theories of similarity as a theoretical foundation for the field of schema matching, critical to the success of BI applications. To this effect, we have reviewed theories with strong credibility in the cognitive research community and have shown how they can be applied to the schema matching context. Importantly, unlike ad-hoc heuristics, a theoretical foundation can provide research directions to the field of schema matching and thereby to the wider area of BI research.

Despite the absence of a single generally accepted theory of similarity, transformational and structural accounts of similarity are highly applicable to the process of schema matching. Both transformational and structural theories operate with constructs pertinent to schema matching. The major strength of the transformational theory is in thoroughly developed computational model that can be easily adapted using existing ETL tools. Expressing similarity through the transformational complexity achieves semantic abstraction and can facilitate automatic matching.

This paper presented a transformational approach using ETL software commonly used in BI applications. In this approach, the computational complexity of transformations, e.g. between *varchar* and *char; money* and *float*, *int* and *decimal* in a relational database, can indicate similarity of the information contained in the corresponding columns. Further, some database management systems support user-defined data types. For example, *str_SSN* in database A and *Social Security No* in database B can be modelled as user-defined data types based on a *char* type. Conversions between base data types and the rules used to define them can also be useful for determining correspondence between the two columns.

Transformational complexity also appears to offer a mechanism for instance-based integration. Transformations can be used to examine the similarity of a variety of stimuli (see, for example, Barenholtz and Tarr, 2008; Cooper and Podgorny, 1976) and data within columns can be compared using conditional complexity. This may be a useful extension of schema transformations. For example, the measure of similarity based on transformational complexity of schema elements can sometimes misguide integration. The field *Sex* in database B is computationally similar to *str_IsInstructor* in database A: both have similar data types, field sizes and constrains; yet the two fields represent different real world phenomena.

The structural model of similarity offers a rich account of the similarity judgment process that encompasses local and global alignment and offers guidance to deal with inconsistent data. As it is not always possible to find exact one-to-one matches between atomic elements of a database, the overall structure becomes more salient. This is captured in the principle of systematicity, which is operationalized by the global alignment principle of the SME and the node activation process in SIAM.

The structural account appears promising for matching major elements of the schema, such as attributes, tables and relations. Yet, focusing on global consistency, structural theories pay little attention to characteristics of table attributes (data types, formulas, constrains). This is where transformational theory appears to offer much practical guidance. Thus, it seems plausible to use structural theory in conjunction with the transformational for particular aspects of schema matching.

The outcome of database integration is central to ensuring current and consistent access to enterprise data by decision support and business intelligence applications. As organizations evolve, so do the information systems supporing their operations. Thus, we expect database integration to be a continuous business process of high importance. While we have proposed an operationalization of the different theories for schema matching, there are still many open questions. For example: what are permissible transformation steps for the application of the RD approach? What should be a "feature" in feature-based theories? To what extent should elements outside of the database, e.g. application software, programming logic, documentation, be considered when making decisions about the schema similarity? Finally, a possibility of unifying several promising theories of similarity in the context of schema matching can be highly important for the development of a general theory of schema matching. These and many other questions are still unanswered, demonstrating the need for continued research, both conceptual as well as empirical.

REFERENCES

1. Agarwal, R., and Karahanna, E. (2000) Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage, *Mis Quarterly* 24, 4, 665-694.
2. Ashby, F.G., and Perrin, N.A. (1988) Toward a Unified Theory of Similarity and Recognition, *Psychological Review* 95, 1, 124-150.
3. Barenholtz, E., and Tarr, M.J. (2008) Visual judgment of similarity across shape transformations: Evidence for a compositional model of articulated objects, *Acta Psychologica* 128, 2, 331-338.
4. Batini, C., Lenzerini, M., and Navathe, S.B. (1986) A comparative analysis of methodologies for database schema integration, *ACM Comput. Surv.* 18, 4, 323-364.
5. Berlin, J., and Motro, A. "Autoplex: Automated Discovery of Content for Virtual Databases," in: *Proceedings of the 9th International Conference on Cooperative Information Systems*, Springer-Verlag, 2001, pp. 108-122.
6. Bin, H., and Che-Chuan Chang, K. (2006) Automatic Complex Schema Matching Across Web Query Interfaces: A Correlation Mining Approach, *ACM Transactions on Database Systems* 31, 1, 346-395.
7. Chen, Y.-P.P., Promparmote, S., and Maire, F. (2006) MDSM: Microarray database schema matching using the Hungarian method, *Information Sciences* 176, 19, 2771-2790.
8. Christie, S., and Gentner, D. (2010) Where Hypotheses Come From: Learning New Relations by Structural Alignment, *Journal of Cognition and Development* 11, 3, 356-373.
9. Cooper, L.A., and Podgorny, P. (1976) Mental transformations and visual comparison processes: Effects of complexity and similarity, *Journal of Experimental Psychology: Human Perception and Performance* 2, 4, 503-514.
10. Doan, A., and Halevy, A.Y. (2005) Semantic-integration research in the database community - A brief survey, *Ai Magazine* 26, 1, 83-94.
11. Domshlak, C., Gal, A., and Roitman, H. (2007) Rank Aggregation for Automatic Schema Matching, *Knowledge and Data Engineering, IEEE Transactions on* 19, 4, 538-553.
12. Evermann, J. (2008a) An Exploratory Study of Database Integration Processes, *Knowledge and Data Engineering, IEEE Transactions on* 20, 1, 99-115.
13. Evermann, J. (2008b) Theories of meaning in schema matching: A review, *Journal of Database Management* 19, 3, 55-82.
14. Evermann, J. (2009) Theories of meaning in schema matching: An exploratory study, *Information Systems* 34, 1, 28-44.
15. Evermann, J. (2010) Contextual Factors in Database Integration — A Delphi Study, in J. Parsons, M. Saeki, P. Shoval, C. Woo and Y. Wand (Eds.) *Conceptual Modeling – ER 2010,* Springer Berlin / Heidelberg, 2010, 274-287.
16. Falkenhainer, B., Forbus, K.D., and Gentner, D. (1989) The structure-mapping engine: Algorithm and examples, *Artificial Intelligence* 41, 1, 1-63.
17. Fan, W., Lu, H., Madnick, S.E., and Cheung, D. (2001) Discovering and reconciling value conflicts for numerical data integration, *Inf. Syst.* 26, 8, 635-656.
18. Forbus, K.D., Usher, J., and Tomai, E. "Analogical learning of visual/conceptual relationships in sketches," in: *Proceedings of the 20th national conference on Artificial intelligence - Volume 1*, AAAI Press, Pittsburgh, Pennsylvania, 2005, pp. 202-208.
19. Gentner, D. (1983) Structure-Mapping: A Theoretical Framework for Analogy*, *Cognitive Science* 7, 2, 155-170.
20. Gentner, D., and Markman, A.B. (1997) Structure Mapping in Analogy and Similarity, *American Psychologist* 52, 1, 45-56.
21. Gentner, D., Rattermann, M.J., and Forbus, K.D. (1993) The Roles of Similarity in Transfer: Separating Retrievability From Inferential Soundness, *Cognitive Psychology* 25, 4, 524-575.
22. Gentner, D., and Toupin, C. (1986) Systematicity and Surface Similarity in the Development of Analogy, *Cognitive Science* 10, 3, 277-300.

23. Goldstone, R.L., and Medin, D.L. (1994) Time Course of Comparison, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20, 1, 29-50.
24. Hahn, U., Chater, N., and Richardson, L.B. (2003) Similarity as transformation, *Cognition* 87, 1, 1.
25. Hahn, U., Close, J., and Graf, M. (2009) Transformation direction influences shape-similarity judgments: Research Article, *Psychological Science* 20, 4, 447-454.
26. Hodgetts, C.J., Hahn, U., and Chater, N. (2009) Transformation and alignment in similarity, *Cognition* 113, 1, 62-79.
27. Holyoak, K.J., and Thagard, P. (1989) Analogical Mapping by Constraint Satisfaction, *Cognitive Science* 13, 3, 295-355.
28. Imai, S. (1977) Pattern similarity and cognitive transformations, *Acta Psychologica* 41, 6, 433-447.
29. Imhoff, C. (2005) Understanding the Three E's of Integration, *DM Review* 15, 4, 8-64.
30. Kang, J., and Naughton, J.F. "On schema matching with opaque column names and data values," in: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, ACM, San Diego, California, 2003, pp. 205-216.
31. Larkey, L.B., and Markman, A.B. (2005) Processes of Similarity Judgment, *Cognitive Science* 29, 6, 1061-1076.
32. Lovett, A., Gentner, D., Forbus, K., and Sagi, E. (2009a) Using analogical mapping to simulate time-course phenomena in perceptual similarity, *Cognitive Systems Research* 10, 3, 216-228.
33. Lovett, A., Tomai, E., Forbus, K., and Usher, J. (2009b) Solving Geometric Analogy Problems Through Two-Stage Analogical Mapping, *Cognitive Science* 33, 7, 1192-1231.
34. Markman, A.B., and Gentner, D. (1993) Splitting the Differences: A Structural Alignment View of Similarity, *Journal of Memory and Language* 32, 4, 517-535.
35. Miller, R.J., Hernandez, M.A., Haas, L.M., Yan, L., Ho, C.T.H., Fagin, R., and Popa, L. (2001) The Clio project: managing heterogeneity, *SIGMOD Rec.* 30, 1, 78-83.
36. Parsons, J., and Wand, Y. (2000) Emancipating Instances from the Tyranny of Classes in Information Modeling, *ACM Transactions on Database Systems* 25, 2, 228.
37. Rahm, E., and Bernstein, P.A. (2001) A survey of approaches to automatic schema matching, *The VLDB Journal* 10, 4, 334-350.
38. Rosch, E. (1975) Basic Objects in Natural Categories, *Bulletin of the Psychonomic Society* 6, Nb4, 415-415.
39. Rosch, E. (1978) Principles of Categorization, in E. Rosch and B. Lloyd (Eds.) *Cognition and Categorization,* John Wiley & Sons Inc, 1978, 27-48.
40. Rumelhart, D.E., and Abrahamson, A.A. (1973) A model for analogical reasoning, *Cognitive Psychology* 5, 1, 1-28.
41. Santini, S., and Jain, R. (1999) Similarity Measures, *IEEE Trans. Pattern Anal. Mach. Intell.* 21, 9, 871-883.
42. Schwering, A. (2005) Hybrid Model for Semantic Similarity Measurement, in R. Meersman and Z. Tari (Eds.) *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE,* Springer Berlin / Heidelberg, 2005, 1449-1465.
43. Shen, R., Vemuri, N.S., Fan, W., and Fox, E.A. (2008) Integration of complex archeology digital libraries: An ETANA-DL experience, *Information Systems* 33, 7-8, 699-723.
44. Shepard, R. (1962) The analysis of proximities: Multidimensional scaling with an unknown distance function. I, *Psychometrika* 27, 2, 125-140.
45. Simitsis, A., and Vassiliadis, P. (2008) A method for the mapping of conceptual designs to logical blueprints for ETL processes, *Decision Support Systems* 45, 1, 22-40.
46. Sjöberg, L. (1969) The Dimensionality of Similarity Judgments, *The American Journal of Psychology* 82, 4, 441-456.
47. Spellman, B.A., and Holyoak, K.J. (1992) If Saddam Is Hitler Then Who Is George Bush? Analogical Mapping Between Systems of Social Roles, *Journal of Personality and Social Psychology* 62, 6, 913-933.
48. Tversky, A. (1977) Features of similarity, *Psychological Review* 84, 4, 327-352.
49. Yi, G., Xiaolong, W., and Qiang, W. (2008) A New Measurement of Systematic Similarity, *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 38, 4, 743-758.
50. Zhao, H., and Ram, S. (2007) Combining schema and instance information for integrating heterogeneous data sources, *Data Knowl. Eng.* 61, 2, 281-303.