Automatch Revisited

Amihai Motro

Abstract We revisit the Autoplex and Automatch projects from 2001–2005, and in particular the results reported in the paper *Database Schema Matching Using Machine Learning with Feature Selection*, presented in the 14th International Conference on Advanced Information Systems Engineering (2002). We provide the motivation and background for these projects, examine their impact a decade later, and sketch possible research directions.

1 Virtual Databases

The problem of integrating information from multiple, independent and heterogeneous information sources has received considerable attention for almost four decades. The problem has been addressed in different ways [2], but the general approach has been that of creating a *virtual database*, a term originally suggested in [8]. In this approach, illustrated in Fig. 1, a single database schema is created that provides an integrated view of the information included in all the information sources. But whereas a conventional database includes data that correspond to its schema, in a virtual database this content is replaced with a *mapping* that associates the elements of this "global" schema with the corresponding data in the component databases. Users may then present queries to the global schema. Using the mapping, each such query is decomposed by the database system into a set of queries that are presented to the component databases; the answers retrieved are then assembled in a single answer that is returned to the user. Ideally, this entire process should be transparent: that is, users should be unaware that the database they are querying is virtual. The main advantage of this approach is that it can be applied in situations in which *physical* integration (i.e., the construction of a new database that incorporates

A. Motro (✉)
Computer Science Department, George Mason University, Fairfax, VA 22030, USA
e-mail: ami@gmu.edu

J. Bubenko et al. (eds.), *Seminal Contributions to Information Systems Engineering*,
DOI 10.1007/978-3-642-36926-1_26,

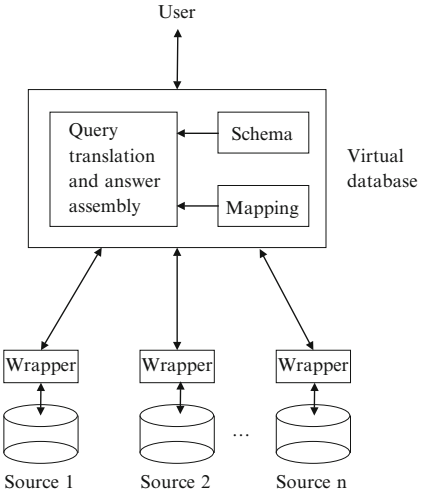

Fig. 1 The architecture of a virtual database system

the information in all the component databases) is impractical—either because it would be too costly or too time-consuming.

This area was continuously evolving during the 1980s, and then enjoyed substantial increase of interest in the 1990s with the decision of DARPA to sponsor research projects in this area. Projects from that period include TSIMMIS [6], Multiplex [10] and Clio [7].

Whereas the original problem assumed mostly a small and static scenario—a small number of databases were to be integrated, and the virtual solution was to endure for an extended period—with the explosion of the Internet in the late 1990s, this scenario has changed dramatically. The set of databases could now be much larger and more dynamic: New relevant databases could become available frequently and would need to be integrated, and old databases might become unavailable and would need to be withdrawn.

Additionally, it has been recognized that the main contributor to the cost of constructing virtual databases was the creation and maintenance of the global schema. This task required comprehending the schemas of the component databases and matching them to the global schema—a laborious task. And when this matching has to be extended and adjusted with each new source, it becomes prohibitively expensive. In brief, a scale-up of the problem needed to be addressed.

2 The Challenge

Given this scenario, it became obvious that one should attempt to automate, at least in part, the process of mapping the global schema to the schemas of the component databases. The intuitive metaphor imagined that one picks up a sheet of paper from the floor and sees a table of some sort; the table is just a grid of values, without any explanations and even without column headings. One then

wonders: "What is the meaning of these values?", and "Can I incorporate them into my accumulated knowledge?" In Internet terms, we envisioned a large collection of sites containing tabular data, possibly discovered with the help of a search engine that was presented with a query comprising global schema terms (e.g., column headings). Each table would then be analyzed to uncover its semantics, and an attempt would be made to match it to the global schema. (This would not only incorporate sources automatically, but would *discover* new sources and thus enrich the virtual database.)

The project was named Autoplex[1] and was described in [3]. It was then observed that a critical part of this automatic resource discovery and global-local schema mapping can be abstracted and generalized to address additional problem domains. Specifically, the problem of matching two independent database schemas (not necessarily global vs. local) has applications, among others, in data warehousing and e-commerce [11]. This sub-project was named Automatch and was the subject of the paper being revisited here.

Automatch applies techniques from machine learning. It assumes a knowledge-base about schema attributes, which has been constructed from examples (in the case of the Autoplex application, these examples would be mappings of the global schema to several component schemas, to be performed manually by experts). This knowledge-base is utilized whenever two new schemas need to be matched. Simply put, the knowledge-base helps score every possible matching of an attribute from one schema with an attribute of the other schema (and eventually, every comprehensive matching of the complete set of attributes of one schema with the complete set of attributes of the other schema).

3 Impact and Future

The work has been well-received and has enjoyed a fair number of bibliographic citations. Arguably, this success is due to the fact that it is part of a trend in the field of information systems to accept solutions that are good but not necessarily perfect. In the pre-Internet years, the focus has been on "critical" applications. Typical domains of interest would be financial, engineering or defense, and there had to be absolute certainty that all solutions are perfect: that data in the database are at all times consistent with the real world, that every answer to a query is precise, and that two databases are integrated flawlessly. While there has been on-going work on information uncertainty and "soft" solutions to information systems tasks (e.g., [1, 9]), with the possible exception of information retrieval, such work has been outside the mainstream, and was generally not awarded center-stage status. The magnitude of the information made available on the Internet has convinced the research community that good but imperfect solutions are sometimes the only available recourse.

[1] The name was a reference to earlier virtual database projects called Multiplex and Fusionplex; Retroplex would follow...

Schema matching and database integration deal with data that are *structured*, typically in tables. With most of the information in the public domain being unstructured, the new challenges are to identify correspondence and similarity between information items that are not values in tables (possibly phrases of free text), and to virtually aggregate and integrate non-structured repositories of information. Possibly, this could be approached by imposing some type of structure on the unstructured information items. A related effort was described in [5]. Indeed, information might be encapsulated in *services*, where stored functions deliver information in response to query-like requests. An attempt to find correspondence among such information items and cluster them in repositories was recently described in [4].

Acknowledgements The original paper was co-authored by Jacob Berlin, who was my doctoral student at that time. Jake deserves an equal share of the credit for the work that we have accomplished. Unfortunately, I was unable to contact Jake for the purpose of this article.

References

1. Andreasen, T., Christiansen, H., Larsen, H.L. (Editors). *Flexible Query Answering Systems*. Kluwer Academic Publishers, 1997.
2. Batini, C., Lenzerini, M., Navathe, S.B. A comparative analysis of methodologies for database schema integration. Computing Surveys, 18(4):323–364, 1989.
3. Berlin, J., Motro, A. Autoplex: Automated discovery of contents for virtual databases. In *Proceedings of COOPIS 01, Sixth IFCIS International Conference on Cooperative Information Systems*, Trento, Italy. Lecture Notes in Computer Science No. 2172, pp. 108–122, 1999.
4. Church, J., Motro, A. Learning service behavior with progressive testing. In *Proceedings of SOCA 11, IEEE International Conference on Service-Oriented Computing and Applications*, Irvine, CA, USA. pp. 1–8, 2011.
5. Etzioni, O., Halevy, A., Doan, A., Ives, Z.G, Madhaven, J., McDowell, L., Tatarinov, I. Crossing the structure chasm. In *Proceedings of CIDR-03, First Biennial Conference on Innovative Data Systems Research*, Asilomar, CA, USA., 2003. Available online at http://www-db.cs.wisc.edu/cidr/cidr2003/program/p11.pdf.
6. Garcia-Molina, H., Papakonstantinou, Y., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V., Widom, J. The TSIMMIS approach to mediation: data models and languages. *Journal of Intelligent Information Systems*, 8(2):117–132, 1997.
7. Miller, R.J., Hernandez, M.A., Haas, L.M., Yan, L., Ho, C.T.H., Fagin, R., Popa, L. The Clio project: managing heterogeneity. *SIGMOD Record* 30(1):78–83, 2001.
8. Motro, A. Interrogating superviews. In *Proceedings of ICOD-2, Second International Conference on Databases*, Cambridge, England, pp. 107–126, 1981.
9. Motro, A., Smets, P. *Uncertainty Management in Information Systems: from Needs to Solutions*. Kluwer Academic Publishing, 1996.
10. Motro, A. Multiplex: a formal model for multidatabases and its implementation. In *Proceedings of NGITS 96, Fourth International Workshop on Next Generation Information Tecnologies and Systems*, Zichron Yaacov, Israel. Lecture Notes in Computer Science No. 1649, pp. 138–158, 1999.
11. Rahm, E., Bernstein, P.A. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4), pp. 334–350, 2001.