Knowledge Discovery su Schemi per l'Integrazione di Basi di Dati

Massimo La Camera[1,†], Luigi Palopoli[2,†], Domenico Saccà[2] and Domenico Ursino[3,*,†]

[1] *Tecnoter s.r.l*

[2] *DIMES, Università della Calabria*

[2] *DII,Università Politecnica delle Marche*

Abstract

In the early 1990s, with the advent of computer networks, the need to integrate heterogeneous databases became increasingly important. This activity was complex but at the same time challenging, since the heterogeneities to be managed were varied and concerned both the extensional component and, perhaps most importantly, the intensional one. Several research groups in Italy and all over the world started to propose solutions to this problem.

1. Introduction: DIKE and XIKE

In the early 1990s, the increasing spread of computer networks opened up new horizons and, at the same time new challenges, in all areas of computer science. The world of databases was no exception, and from the very beginning it was clear that the possibility of integrating heterogeneous databases and making them work together would have enormous benefits, but also pose challenging issues. In those years, the "reigning" logic model was the relational one. Therefore, the heterogeneity was not so much in the data representation model as in the data itself and the conventions used for it (for instance, the string "BL" could represent the color blue in one database and the color black in another). In addition to the heterogenity of extensional component, there was the heterogeneity of intensional one (i.e., regarding schemas and semantics), which was undoubtedly the most difficult to manage. In fact, when integrating different database schemas, it was necessary to detect and handle synonymies (i.e., the same concept represented by different names in different databases), homonymies (i.e., different concepts represented by the same name in different databases), hyponymies/hyperonimies (i.e., the presence in a schema of a concept that is a specialization of another concept from another schema), and so on.

Several research groups in Italy and around the world have taken up this challenge and started to propose solutions. One of them was the group from University of Calabria, which focused

SEBD 2024: The 32nd Italian Symposium on Advanced Database Systems, June 23-26, 2024 - Villasimius, Sardinia, Italy

[*] Corresponding author.

[†] These authors contributed equally.

massimo.lacamera@tecnoter.net (M. La Camera); palopoli@dimes.unical.it (L. Palopoli); sacca@dimes.unical.it (D. Saccà); d.ursino@univpm.it (D. Ursino)

0000-0003-4915-5137 (L. Palopoli); 0000-0003-3584-5372 (D. Saccà); 0000-0003-1360-8499 (D. Ursino)

mainly on extracting interschema properties (i.e., synonymies, homonymies, hyponymies/hyperonimies, subschema similarities) and using them for database integration. The first solution proposed by this group was presented at SEBD in 1997 [1]. This solution, after several studies and refinements described in papers published at SEBD [2, 3, 4], as well as in conference proceedings [5, 6, 7, 8, 9] and in prestigious international journals [10, 11, 12, 13, 14], gave rise to the DIKE (Database Intensional Knowledge Extractor) system [15, 16].

Underlying the DIKE approach was a seemingly very simple, yet powerful concept that is present in various forms in many areas of computer science and research in general. In fact, DIKE assumed that, given two concepts from different databases, if their "neighborhing" concepts in the databases they belong to were similar, then they were probably similar; conversely, if their "neighborhing" concepts were different, then they were probably different.

We feel it necessary to point out that in Italy, in the same years, the research group of the University of Brescia and the University of Milan, which proposed ARTEMIS [17, 18, 19], and the research group of the University of Modena and Reggio Emilia, which proposed MOMIS [20, 21], were working on the same issues. Abroad, several research groups strove to address the same challenge. Among them, one of the most renowned was undoubtedly the group of Prof. Philip Bernstein at University of Seattle and Microsoft Research, which proposed Cupid [22].

An important acknowledgement to the database integration Italian school came from Prof. Bernstein himself, who in a famous paper [22] proposed a detailed comparison between Cupid, DIKE and ARTEMIS/MOMIS, recognizing how the latter two systems were able to "compete on a par" with Cupid. Again DIKE and ARTEMIS/MOMIS were considered, along with Cupid, in another important paper [23], which proposed an account of schema matching research ten years after the publication of [22].

Over the years, the relational model, while still extremely important in the database world, showed all of its limitations when dealing with semi-structured and unstructured data. To handle semi-structured data, the Object Exchange Model (OEM) and later the Extensible Markup Language (XML) and JavaScript Object Notation (JSON) were proposed over the years. The authors of DIKE continued their research in this area and presented approaches capable of extracting and handling interschema properties from semi-structured data [24, 25, 26, 27]. This research effort led over the years to the XIKE (XML source Intensional Knowledge Extractor) system [28].

The SEBD Community has followed this stream of innovation, first with the "progenitors" of DIKE [1], and later with DIKE [2, 3, 4] and XIKE [25, 27]. Research on DIKE and XIKE has also received some awards. These include the Best Student Paper Award at the International Symposium on Advances in Databases and Information Systems (ADBIS'99) [29] and the publication of Domenico Ursino's PhD thesis in Springer's Lecture Notes in Computer Science series [30]. Both DIKE and XIKE have been acquired by companies for use as cores within frameworks aimed at managing Cooperative Information Systems.

DIKE and XIKE represented the apex of the studies of the database integration group of the University of Calabria. By the middle of the first decade of the new century, the members of the group were interested in new topics such as intelligent agents, recommender systems, data mining, social network analysis, and deep learning. However, these members were able to observe how some of the ideas that were the basis of DIKE/XIKE were adopted, perhaps under different forms and names, in research in the areas they were directly concerned with and in others. To take just one example, the idea behind DIKE/XIKE that the semantics of a concept depends on its neighbors is used in collaborative filtering recommender systems, when we say that the interests of a person are influenced by the ones of people closest to her/him, and is the basis of the concept of homophily [31] in social network analysis. A researcher should not be surprised that there are some ideas/principles so powerful and general that they can be used successfully in a variety of fields. However, discovering this through direct personal experience is always astonishing, despite the many years of research she/he may have behind her/him.

References

[1] M. L. Camera, L. Palopoli, D. Saccà, D. Ursino, Knowledge discovery su schemi per l'integrazione di sistemi di basi di dati, in: Atti del Congresso sui Sistemi Evoluti per Basi di Dati (SEBD'97), Verona, Italy, 1997, pp. 166–190. In Italian.

[2] A. Bonifati, L. Palopoli, D. Saccà, D. Ursino, Utilizzo della logica descrittiva per l'estrazione di proprietà terminologiche e strutturali complesse, in: Atti del Congresso su Sistemi Evoluti per Basi di Dati (SEBD'98), Ancona, Italy, 1998, pp. 71–86. In Italian.

[3] L. Palopoli, L. Pontieri, D. Ursino, Progettazione semi-automatica di data warehouse di grandi dimensioni, in: Atti del Congresso su Sistemi Evoluti per Basi di Dati (SEBD'99), Como, Italy, 1999, pp. 3–17. In Italian.

[4] L. Palopoli, G. Terracina, D. Ursino, Derivazione di iponimie/iperonimie tra entità appartenenti a basi di dati eterogenee, in: Atti del Congresso su Sistemi Evoluti per Basi di Dati (SEBD 2000), L'Aquila, Italy, 2000, pp. 357–370. In Italian.

[5] L. Palopoli, D. Saccà, D. Ursino, Semi-automatic, semantic discovery of properties from database schemes, in: Proc. of the International Database Engineering and Applications Symposium (IDEAS '98), Cardiff (Wales), UK, 1998, pp. 244–253. IEEE Computer Society.

[6] L. Palopoli, D. Saccà, D. Ursino, An automatic technique for detecting type conflicts in database schemes, in: Proc. of the ACM International Conference on Information and Knowledge Management (CIKM'98), Bethesda, Maryland, USA, 1998, pp. 306–313. ACM Press.

[7] D. Ursino, Deriving type conflicts and object cluster similarities in database schemes by an automatic and semantic approach, in: Proc. of the International Symposium on Advances in Databases and Information Systems (ADBIS'99), Maribor, Slovenia, 1999, pp. 46–60. Lecture Notes in Computer Science, Springer-Verlag.

[8] L. Palopoli, D. Saccà, G. Terracina, D. Ursino, A unified graph-based framework for deriving nominal interscheme properties, type conflicts and object cluster similarities, in: Proc. of the International Conference on Cooperative Information Systems (CoopIS'99), Edinburgh, Scotland, United Kingdom, 1999, pp. 34–45. IEEE Computer Society.

[9] G. Terracina, D. Ursino, A study on the interaction between interscheme property extraction and type conflict resolution, in: Proc. of the International Database Engineering and Applications Symposium (IDEAS '00), Yokohama, Japan, 2000, pp. 25–33. IEEE Computer Society.

[10] L. Palopoli, D. Saccà, D. Ursino, Semi-automatic techniques for deriving interscheme properties from database schemes, Data & Knowledge Engineering 30(4) (1999) 239–273.

[11] L. Palopoli, D. Saccà, D. Ursino, DL$_{\mathcal{P}}$: a description logic for extracting and managing complex terminological and structural properties from database schemes, Information Systems 24(5) (1999) 410–424.

[12] L. Palopoli, L. Pontieri, G. Terracina, D. Ursino, Intensional and extensional integration and abstraction of heterogeneous databases., Data & Knowledge Engineering 35(3) (2000) 201–237.

[13] G. Terracina, D. Ursino, A uniform methodology for extracting type conflicts and subscheme similarities from heterogeneous databases, Information Systems 25(8) (2000) 527–552.

[14] L. Palopoli, D. Saccà, G. Terracina, D. Ursino, Uniform techniques for deriving similarities of objects and subschemes in heterogeneous databases, IEEE Transactions on Knowledge and Data Engineering 15(2) (2003) 271–294.

[15] L. Palopoli, G. Terracina, D. Ursino, Experiences using DIKE, a system for supporting cooperative information system and data warehouse design, Information Systems 28(7) (2003) 835–865.

[16] L. Palopoli, G. Terracina, D. Ursino, DIKE: a system supporting the semi-automatic construction of Cooperative Information Systems from heterogeneous databases, Software Practice & Experience 33(9) (2003) 847–884.

[17] S. Castano, V. D. Antonellis, Reference conceptual architecture for re-engineering information systems, International Journal of Cooperative Information Systems 4(2) (1995) 213–235.

[18] S. Castano, V. D. Antonellis, M. Fugini, B. Pernici, Conceptual schema analysis: Techniques and applications, ACM Transactions on Database Systems (TODS) 23 (1998) 286–332.

[19] S. Castano, V. D. Antonellis, Building views over semistructured data sources, in: Proc. of the International Conference on Conceptual Modeling (ER'99), Paris, France, 1999, pp. 146–160. Springer.

[20] S. Bergamaschi, S. Castano, M. Vincini, Semantic integration of semistructured and structured data sources, SIGMOD Record 28(1) (1999) 54–59.

[21] S. Bergamaschi, S. Castano, M. Vincini, D. Beneventano, Semantic integration and query of heterogeneous information sources, Data & Knowledge Engineering 36(3) (2001) 215–249.

[22] J. Madhavan, P. Bernstein, E. Rahm, Generic schema matching with Cupid, in: Proc. of the International Conference on Very Large Data Bases (VLDB 2001), Rome, Italy, 2001, pp. 49–58. Morgan Kaufmann.

[23] P. Bernstein, J. Madhavan, E. Rahm, Generic Schema Matching, Ten Years Later, Proceedings of the VLDB Endowment 4 (2011) 695–701.

[24] G. Terracina, D. Ursino, Deriving synonymies and homonymies of object classes in semi-structured information sources, in: Proc. of the International Conference on Management of Data (COMAD 2000), Pune, India, 2000, pp. 21–32. McGraw Hill.

[25] P. De Meo, G. Quattrone, G. Terracina, D. Ursino, Estrazione, a vari livelli di “severità”, di proprietà interschema da Schemi XML, in: Atti del Congresso sui Sistemi Evoluti per Basi di Dati (SEBD 2004), S. Margherita di Pula (Cagliari), Italy, 2004, pp. 290–301.

[26] P. De Meo, G. Quattrone, G. Terracina, D. Ursino, Integration of XML Schemas at various

"severity" levels, Information Systems 31(6) (2006) 397–434.

[27] P. De Meo, G. Quattrone, G. Terracina, D. Ursino, Utilizzo delle proprietà interschema per il clustering di schemi XML semanticamente eterogenei, in: Atti del Congresso sui Sistemi Evoluti per Basi di Dati (SEBD 2005), Brixen-Bressanone, Italy, 2005, pp. 336–347. Aracne.

[28] P. De Meo, G. Quattrone, G. Terracina, D. Ursino, Experiences with the system XIKE (XML source Intensional Knowledge Extractor), Soft Computing: New Research (2009) 333–386. Nova Science notes.

[29] L. Palopoli, G. Terracina, D. Ursino, The system dike: Towards the semi-automatic synthesis of cooperative information systems and data warehouses, in: Proc. of the Challenges of Symposium on Advances in Databases and Information Systems (ADBIS-DASFAA 2000), Prague, Czech Republic, 2000, pp. 108–117. Matfyzpress.

[30] D. Ursino, Extraction and Exploitation of Intensional Knowledge from Heterogeneous Information Sources, Heidelberg, Germany, 2002. PhD Thesis, Lecture Notes in Computer Science 2282, Springer Verlag.

[31] M. McPherson, L. Smith-Lovin, J. Cook, Birds of a feather: Homophily in social networks, Annual Review of Sociology 27 (2001) 415–444. JSTOR.