Bibliography

[1] *The Description Logic Handbook: Theory, Implementation and Applications, 2nd Edition*. Cambridge University Press, 2007.

[2] 2011.

[3] 2011.

[4] 2011.

[5] Daniel J. Abadi. Tradeoffs between parallel database systems, hadoop, and hadoopdb as platforms for petabyte-scale analysis. In *SSDBM*, pages 1–3, 2010.

[6] Karl Aberer, Philippe Cudre-Mauroux, and Manfred Hauswirth. The chatty web: Emergent semantics through gossiping. In *Proc. of the Int. WWW Conf.*, 2003.

[7] Serge Abiteboul, Omar Benjelloun, Bogdan Cauytis, Ioana Manolescu, Tova Milo, and Nicoleta Preda. Lazy query evaluation for Active XML. In *SIGMOD 2004, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 13-18, 2004, Paris, France*, June 2004.

[8] Serge Abiteboul, Angela Bonifati, Gregory Cobena, Ioana Manolescu, and Tova Milo. Dynamic XML documents with distribution and replication. In *SIGMOD 2003, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 9-12, 2003, San Diego, California, USA*, June 2003.

[9] Serge Abiteboul and Oliver M. Duschka. Complexity of Answering Queries Using Materialized Views. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, 1998.

[10] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison Weseley, 1995.

[11] B. Adelberg. Nodose - a tool for semi-automatically extracting semi-structured data from text documents. In *SIGMOD*, 1998.

[12] P. Adjiman, Philippe Chatalic, François Goasdoué, Marie-Christine Rousset, and Laurent Simon. Distributed reasoning in a peer-to-peer setting. In *ECAI*, pages 945–946, 2004.

[13] Philippe Adjiman, François Goasdoué, and Marie-Christine Rousset. Somerdfsin the semantic web. *J. Data Semantics*, 8:158–181, 2007.

[14] Foto Afrati, Chen Li, and Prasenjit Mitra. Rewriting queries using views in the presence of arithmetic comparisons. *Theoretical Computer Science*, 368(1-2):88–123, 2006.

[15] Charu Aggrawal, editor. *Managing and Mining Uncertain Data*. Kluwer Academic Publishers, 2009.

[16] Sanjay Agrawal, Surajit Chaudhuri, and Gautam Das. Dbxplorer: A system for keyword-based search over relational databases. In *Proc. of ICDE*, pages 5–16, 2002.

[17] Sanjay Agrawal, Surajit Chaudhuri, and Vivek R. Narasayya. Automated selection of materialized views and indexes in SQL databases. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pages 496–505, 2000.

[18] Sanjay Agrawal, Surajit Chaudhuri, and vivek R. Narasayya. Automated selection of materialized views and indexes in SQL databases. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*. Morgan Kaufman, 2000.

[19] Rafi Ahmed, Phillippe De Smedt adn Weimin Du, William Kent, Mohammad A. Ketabchi, Witold A. Litwin, Abbas Rafii, and Ming-Chien Shan. The pegasus heterogeneous multidatabase system. *IEEE Computer*, pages 19–26, December 1991.

[20] Shurug Al-Khalifa, H. V. Jagadish, Nick Koudas, Divesh Srivastava, and Yuqing Wu. Structural joins: A primitive for efficient XML query pattern matching. In *Proceedings of the 18th International Conference on Data Engineering, February 26-March 1, 2002, San Jose, CA USA*. IEEE Computer Society, 2002.

[21] Mehmet Altinel and Michael J. Franklin. Efficient filtering of XML documents for selective dissemination of information. In *VLDB 2000, Proceedings of 26th*

International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt. Morgan Kaufman, 2000.

[22] Peter Alvaro, Tyson Condie, Neil Conway, Khaled Elmeleegy, Joseph M. Hellerstein, and Russell Sears. Boom analytics: exploring data-centric, declarative programming for the cloud. In *EuroSys*, pages 223–236, 2010.

[23] Yael Amsterdamer, Daniel Deutch, and Val Tannen. Provenance for aggregate queries. In *PODS*, pages 153–164, 2011.

[24] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *VLDB*, 2002.

[25] Rohit Ananthakrishna, Surajit Chaudhuri, and Venkatesh Ganti. Eliminating Fuzzy Duplicates in Data Warehouses. In *Proc. of VLDB*, 2002.

[26] A. Arasu and H. Garcia-Molina. Extracting structured data from web pages. In *SIGMOD*, 2003.

[27] Arvind Arasu, Venkatesh Ganti, and Raghav Kaushik. Efficient exact set-similarity joins. In *VLDB*, pages 918–929, 2006.

[28] Yigal Arens, Chin Y. Chee, Chun-Nan Hsu, and Craig A. Knoblock. Retrieving and integrating data from multiple information sources. *International Journal on Intelligent and Cooperative Information Systems*, 1994.

[29] Yigal Arens, Craig A. Knoblock, and Wei-Min Shen. Query reformulation for dynamic information integration. *International Journal on Intelligent and Cooperative Information Systems*, (6) 2/3:99–130, June 1996.

[30] G. O. Arocena and A. O. Mendelzon. Weboql: Restructuring documents, databases, and webs. In *ICDE*, 1998.

[31] P. Atzeni and R. Torlone. Management of multiple models in an extensible database design tool. In *Proc. of EDBT*, pages 79–95, 1996.

[32] Paolo Atzeni, Giansalvatore Mecca, and Paolo Merialdo. To weave the web. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, 1997.

[33] Ron Avnur and Joseph M. Hellerstein. Eddies: Continuously adaptive query processing. In *Proc. of SIGMOD*, 2000.

[34] Brian Babcock, Shivnath Babu, Mayur Datar, and Rajeev Motwani. Chain: Operator scheduling for memory minimization in data stream systems. In *SIGMOD 2003, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 9-12, 2003, San Diego, California, USA*, pages 253–264, 2003.

[35] Brian Babcock and Surajit Chaudhuri. Towards a robust query optimizer: a principled and practical approach. In *SIGMOD 2005, Proceedings of the ACM International Conference on Management of Data, June 14-16, 2005, Baltimore, MD*, pages 119–130, New York, NY, USA, 2005. ACM Press.

[36] Brian Babcock, Mayur Datar, and Rajeev Motwani. Load shedding for aggregation queries over data streams. In *Proceedings of the 20th International Conference on Data Engineering, March 30-April 2, 2004, Boston, MA*, pages 350–361, 2004.

[37] Shivnath Babu and Pedro Bizarro. Adaptive query processing in the looking glass. In *CIDR 2005: Second Biennial Conference on Innovative Data Systems Research, Asilomar, CA*, pages 238–249, 2005.

[38] Shivnath Babu, Rajeev Motwani, Kamesh Munagala, Itaru Nishizawa, and Jennifer Widom. Adaptive ordering of pipelined stream filters. In *SIGMOD 2004, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 13-18, 2004, Paris, France*, 2004.

[39] Shivnath Babu, Utkarsh Srivastava, and Jennifer Widom. Exploiting k-constraints to reduce memory overhead in continuous queries over streams. Technical report, Stanford University, 2002.

[40] François Bancilhon and Nicolas Spyratos. Update semantics of relational views. *ACM Transactions on Database Systems*, 6(4):557–575, 1981.

[41] Luciano Barbosa and Juliana Freire. Siphoning hidden-web data through keyword-based interfaces. In *SBBD*, 2004.

[42] Greg Barish and Craig A. Knoblock. Learning value predictors for the speculative execution of information gathering plans. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3–9, 2003.

[43] R. Baumgartner, S. Flesca, and G. Gottlob. Declarative information extraction, Web crawling, and recursive wrapping with Lixto. In *Proc. of the 6th Int Conf. on Logic Programming and Nonmonotonic Reasoning*, 2001.

[44] R. Baumgartner, S. Flesca, and G. Gottlob. Visual Web information extraction with lixto. In *VLDB*, 2001.

[45] Louis Bavoil, Steven P. Callahan, Patricia J. Crossno, Juliana Freire, Carlos E. Scheidegger, Claudio T. Silva, and Huy T. Vo. VisTrails: Enabling interactive multiple-view visualizations. *IEEE Visualization*, 2005.

[46] Roberto J. Bayardo, Yiming Ma, and Ramakrishnan Srikant. Scaling up all pairs similarity search. In *WWW*, pages 131–140, 2007.

[47] Catriel Beeri, Alon Y. Levy, and Marie-Christine Rousset. Rewriting queries using views in description logics. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 99–108, Tucson, Arizona., 1997.

[48] Omar Benjelloun, Anish Das Sarma, Alon Y. Halevy, and Jennifer Widom. ULDBs: Databases with uncertainty and lineage. In *VLDB 2006, Proceedings of 31st International Conference on Very Large Data Bases, September 12-15, 2006, Seoul, Korea*, pages 953–964, 2006.

[49] Sonia Bergamaschi, Silvana Castano, and Maurizio Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record*, 28(1):54–59, 1999.

[50] Michael K. Bergman. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 2001.

[51] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 279, 2001.

[52] P. Bernstein, F. Giunchiglia, A. Kementsietsidis, J. Mylopoulos, L. Serafini, and I. Zaihrayeu. Data management for peer-to-peer computing : A vision. In *Proceedings of the WebDB Workshop*, 2002.

[53] Philip A. Bernstein. Applying model management to classical meta-data problems. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*, 2003.

[54] Philip A. Bernstein and Dah-Ming W. Chiu. Using semi-joins to solve relational queries. *J. ACM*, 28:25–40, 1981.

[55] Philip A. Bernstein, Todd J. Green, Sergey Melnik, and Alan Nash. Implementing mapping composition. In *Proc. of VLDB*, pages 55–66, 2006.

[56] Philip A. Bernstein, Alon Y. Halevy, and Rachel Pottinger. A vision for management of complex models. *SIGMOD Record*, 29(4):55–63, 2000.

[57] Philip A. Bernstein and Sergey Melnik. Model management 2.0: Manipulating richer mappings. In *Proc. of SIGMOD*, pages 1–12, 2007.

[58] Gaurav Bhalotia, Arvind Hulgeri, Charuta Nakhe, Soumen Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In *Proc. of ICDE*, pages 431–440, 2002.

[59] I. Bhattacharya and L. Getoor. A latent Dirichlet model for unsupervised entity resolution. In *Proc. of the SIAM Int. Conf. on Data Mining (SDM)*, 2006.

[60] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[61] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proc. of the ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 39–48, 2003.

[62] M. Bilenko, R. J. Mooney, W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, 2003.

[63] M. Bilenko, R. J. Mooney, W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, 2003.

[64] Olivier Biton, Sarah Cohen Boulakia, Susan B. Davidson, and Carmem S. Hara. Querying and managing provenance through user views in scientific workflows. In *Proceedings of the 24th International Conference on Data Engineering*, 2008.

[65] J. A. Blakeley, N. Coburn, and P. A. Larson. Updating derived relations: detecting irrelevant and autonomously computable updates. *ACM Transactions of Database Systems*, 14(3):369–400, 1989.

[66] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *CACM*, 13(7):422–426, July 1970.

[67] Lukas Blunschi, Jens-Peter Dittrich, Olivier Girard, Shant Krakos Karakashian, and Marcos Antonio Vas Salles. The imemex personal dataspace management system (demo). In *CIDR*, 2007.

[68] Alex Borgida. Description logics in data management. *IEEE Trans. on Know. and Data Engineering*, 7(5):671–682, 1995.

[69] V. R. Borkar, K. Deshmukh, and S. Sarawagi. Automatic segmentation of text into structured records. In *SIGMOD*, 2001.

[70] Michael Brundage. *XQuery: The XML Query Language*. Addison-Wesley Professional, February 2004.

[71] Nicolas Bruno, Nick Koudas, and Divesh Srivastava. Holistic twig joins: optimal xml pattern matching. In *SIGMOD Conference*, pages 310–321, 2002.

[72] Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D. Ernst. HaLoop: Efficient iterative data processing on large clusters. In *36th International Conference on Very Large Data Bases*, Singapore, September 14–16, 2010.

[73] Peter Buneman, Anthony, Susan Davidson, and Kosky. Theoretical aspects of schema merging. In *Proc. of EDBT*, pages 152–167, 1992.

[74] Peter Buneman, Sanjeev Khanna, and Wang Chiew Tan. Why and where: A characterization of data provenance. In Jan Van den Bussche and Victor Vianu, editors, *Database Theory — ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*, pages 316–330. Springer, 2001.

[75] M. J. Cafarella, A. Y. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. WebTables: exploring the power of tables on the Web. *PVLDB*, 1(1):538–549, 2008.

[76] Michael J. Cafarella, Alon Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu. Uncovering the Relational Web. In *WebDB*, 2008.

[77] Michael J. Cafarella, Alon Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu. WebTables: Exploring the Power of Tables on the Web. In *VLDB*, 2008.

[78] D. Cai, S. Yu, J. Wen, and W. Ma. Extracting content structure for Web pages based on visual representation. In *Proc. of the 5th Asian-Pacific Web Conference (APWeb)*, 2003.

[79] Andrea Cali, Diego Calvanese, and Giuseppe DeGiacomo adn Maurizio Lenzerini. Data integration under integrity constraints. In *Proceedings of CAiSE*, pages 262–279, 2002.

[80] M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction. In *AAAI*, 1999.

[81] James P. Callan and Margaret E. Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19(2):97–130, 2001.

[82] D. Calvanese, G. De Giacomo, and M. Lenzerini. Answering queries using views over description logics. In *Proceedings of AAAI*, pages 386–391, 2000.

[83] Michael J. Carey and Donald Kossmann. On saying "enough already!" in sql. In *SIGMOD 1997, Proceedings of the ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 219–230, 1997.

[84] S. Castano and V. De Antonellis. A discovery-based approach to database ontology design. *Distributed and Parallel Databases - Special Issue on Ontologies and Databases*, 7(1), 1999.

[85] T. Catarci and M. Lenzerini. Representing and using interschema knowledge in cooperative information systems. *Journal of Intelligent and Cooperative Information Systems*, pages 55–62, 1993.

[86] Xiaoyong Chai, Ba-Quy Vuong, AnHai Doan, and Jeffrey F. Naughton. Efficiently incorporating user feedback into information extraction and integration programs. In *SIGMOD 2009, Proceedings of the ACM International Conference on Management of Data, June 29-July 2, 2009, Providence, Rhode Island*, pages 87–100, New York, NY, USA, 2009. ACM.

[87] Craig Chambers, Ashish Raniwala, Frances Perry, Stephen Adams, Robert R. Henry, Robert Bradshaw, and Nathan Weizenbaum. FlumeJava: easy, efficient data-parallel pipelines. In *PLDI*, pages 363–375, 2010.

[88] A.K. Chandra and P.M. Merlin. Optimal implementation of conjunctive queries in relational databases. In *Proceedings of the Ninth Annual ACM Symposium on Theory of Computing*, pages 77–90, 1977.

[89] C. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan. A survey of web information extraction systems. *IEEE Trans. Knowl. Data Eng.*, 18(10):1411–1428, 2006.

[90] C. Chang and S. Lui. IEPAD: information extraction based on pattern discovery. In *WWW*, 2001.

[91] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert Gruber. Bigtable: A distributed storage system for structured data (awarded best paper!). In *OSDI*, pages 205–218, 2006.

[92] Adriane Chapman and H. V. Jagadish. Why not? In *SIGMOD Conference*, pages 523–534, 2009.

[93] Sam Chapman. Sam's string metrics. 2006. Available at staffwww.dcs.shef.ac.uk/people/sam.chapman@know.co.uk/stringmetrics.html.

[94] A. Chatterjee and A. Segev. Data manipulation in heterogeneous databases. *SIGMOD Record*, 20(4):64–68, 1991.

[95] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani. Robust and efficient fuzzy match for online data cleaning. In *SIGMOD*, 2003.

[96] Surajit Chaudhuri. An overview of query optimization in relational systems. In *Proceedings of the Seventeenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, June 1-3, 1998, Seattle, Washington, USA*, pages 34–43, 1998.

[97] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehouse and OLAP technology. *SIGMOD Record*, 26(1), March 1997.

[98] Surajit Chaudhuri, Venkatesh Ganti, and Raghav Kaushik. A primitive operator for similarity joins in data cleaning. In *ICDE*, page 5, 2006.

[99] Surajit Chaudhuri and Moshe Vardi. Optimizing real conjunctive queries. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 59–70, Washington D.C., 1993.

[100] Sudarshan Chawathe, Hector Garcia-Molina, Joachim Hammer, Kelly Ireland, Yannis Papakonstantinou, Jeffrey Ullman, and Jennifer Widom. The TSIMMIS project: Integration of heterogeneous information sources. In proceedings of IPSJ, Tokyo, Japan, October 1994.

[101] Chandra Chekuri and Anand Rajaraman. Conjunctive query containment revisited. *Theor. Comput. Sci.*, 239(2):211–229, 2000.

[102] Yi Chen, Susan B. Davidson, and Yifeng Zheng. Vitex: A streaming xpath processing system. In *ICDE*, pages 1118–1119, 2005.

[103] James Cheney, Amal Ahmed, and Umut A. Acar. Provenance as dependency analysis. *Mathematical Structures in Computer Science (MSCS)*, April 2011.

[104] James Cheney, Laura Chiticariu, and Wang Chiew Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474, 2009.

[105] L. Chiticariu, P. G. Kolaitis, and L. Popa. Interactive generation of integrated schemas. In *Proc. of SIGMOD*, 2008.

[106] L. Chiticariu, W. Tan, and G. Vijayvargiya. DBNotes: a post-it system for relational databases based on provenance. *Proc. of ACM SIGMOD*, 2005.

[107] Francis C. Chu, Joseph Y. Halpern, and Johannes Gehrke. Least expected cost query optimization: What can we expect? In *PODS*, pages 293–302, 2002.

[108] S. Chuang, K. C. Chang, and C. Zhai. Collaborative wrapping: A turbo framework for Web data extraction. In *ICDE*, 2007.

[109] M. Cochinwala, V. Kurien, G. Lalk, and D. Shasha. Efficient data reconciliation. *Inf. Sci.*, 137(1-4):1–15, 2001.

[110] Sara Cohen. Containment of aggregate queries. *SIGMOD Record*, 34(1):77–85, 2005.

[111] W. Cohen. A mini-course on record linkage and matching, 2004. `http://www.cs.cmu.edu/~wcohen`.

[112] W. Cohen and J. Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proc. of the ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 475–480, 2002.

[113] W. W. Cohen. Record linkage tutorial: Distance metrics for text. 2001. PPT slides, available at `www.cs.cmu.edu/~wcohen/Matching-2.ppt`.

[114] W. W. Cohen, M. Hurst, and L. S. Jensen. A flexible learning system for wrapping tables and lists in HTML documents. In *WWW*, 2002.

[115] W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, 2003.

[116] William W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Trans. Inf. Syst.*, 18(3):288–321, 2000.

[117] Richard L. Cole and Goetz Graefe. Optimization of dynamic query evaluation plans. In *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, Minneapolis, Minnesota, May 24-27, 1994*, pages 150–160. ACM Press, 1994.

[118] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Sean Slattery. Learning to extract symbolic knowledge from the world-wide web. In *Proceedings of the AAAI Fifteenth National Conference on Artificial Intelligence*, 1998.

[119] V. Crescenzi and G. Mecca. Grammars have exceptions. *Inf. Syst.*, 23(8):539–565, 1998.

[120] V. Crescenzi and G. Mecca. Automatic information extraction from large websites. *J. ACM*, 51(5):731–779, 2004.

[121] V. Crescenzi, G. Mecca, and P. Merialdo. Roadrunner: Towards automatic data extraction from large web sites. In *VLDB*, 2001.

[122] Yingwei Cui. *Lineage Tracing in Data Warehouses*. PhD thesis, Stanford Univ., 2001.

[123] A. Culotta and A. McCallum. Joint deduplication of multiple record types in relational data. In *Proc. of the ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 257–258, 2005.

[124] N. Dalvi, R. Kumar, and M. A. Soliman. Automatic wrappers for large scale web extraction. *PVLDB*, 4(4):219–230, 2011.

[125] N. N. Dalvi, P. Bohannon, and F. Sha. Robust Web extraction: an approach based on a probabilistic tree-edit model. In *SIGMOD*, 2009.

[126] Dean Daniels. Query compilation in a distributed database system. Technical Report RJ 3423, IBM, 1982.

[127] A. Das Sarma, L. Dong, and A. Halevy. Bootstrapping pay-as-you-go data integration systems. In *Proc. of SIGMOD*, 2008.

[128] Susan B. Davidson, Sanjeev Khanna, Tova Milo, Debmalya Panigrahi, and Sudeepa Roy. Provenance views for module privacy. In *PODS*, pages 175–186, 2011.

[129] U. Dayal. Processing queries over generalized hierarchies in a multidatabase systems. In *Proc. of the VLDB Conf.*, pages 342–353, 1983.

[130] Umeshwar Dayal and Philip A. Bernstein. On the correct translation of update operations on relational views. *ACM Transactions on Database Systems*, 7(3):381–416, 1982.

[131] Filipe de S Mesquita, Altigran Soares da Silva, Edleno Silva de Moura, Pvel Calado, and Alberto H. F. Laender. Labrador: Efficiently publishing relational databases on the web by using keyword-based query interfaces. *Inf. Process. Manage.*, 43(4):983–1004, 2007.

[132] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI*, pages 137–150, 2004.

[133] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: amazon's highly available key-value store. *SIGOPS Oper. Syst. Rev.*, 41(6):205–220, 2007.

[134] L. G. DeMichiel. Resolving Database Incompatibility: An Approach to Performing Relational Operations over Mismatched Domains. *IEEE Transactions on Knowledge and Data Engineering*, 1989.

[135] Pedro DeRose, Warren Shen, Fei Chen, AnHai Doan, and Raghu Ramakrishnan. Building structured web community portals: A top-down, compositional, and incremental approach. In *Proc. of VLDB*, 2007.

[136] Amol Deshpande and Joseph M. Hellerstein. Lifting the burden of history from adaptive query processing. In *VLDB 2004, Proceedings of 30th International Conference on Very Large Data Bases, August 29-September 3, 2004, Toronto, Canada*, pages 948–959. Morgan Kaufman, 2004.

[137] Amol Deshpande, Zachary Ives, and Vijayshankar Raman. Adaptive query processing. *Foundations and Trends in Databases*, 2007.

[138] Stefan Dessloch, Mauricio A. Hernández, Ryan Wisnesky, Ahmed Radwan, and Jindan Zhou. Orchid: Integrating schema mapping and etl. In *ICDE*, pages 1307–1316, 2008.

[139] Stefan Dessloch, Mauricio A. Hernández, Ryan Wisnesky, Ahmed Radwan, and Jindan Zhou. Orchid: Integrating schema mapping and etl. In *ICDE*, pages 1307–1316, 2008.

[140] Alin Deutsch, Mary F. Fernández, Daniela Florescu, Alon Y. Levy, and Dan Suciu. XML-QL. In *QL*, 1998.

[141] Alin Deutsch and Val Tannen. MARS: A system for publishing XML from mixed and redundant storage. In *VLDB 2003, Proceedings of 29th International Conference on Very Large Data Bases, September 9-12, 2003, Berlin, Germany.* Morgan Kaufman, 2003.

[142] D. Dey. Entity matching in heterogeneous databases: A logistic regression approach. *Decision Support Systems*, 44(3):740–747, 2008.

[143] D. Dey, S. Sarkar, and P. De. A distance-based approach to entity reconciliation in heterogeneous databases. *IEEE Trans. Knowl. Data Eng.*, 14(3):567–582, 2002.

[144] Hong-Hai Do and Erhard Rahm. COMA - a system for flexible combination of schema matching approaches. In *Proc. of VLDB*, 2002.

[145] AnHai Doan, Pedro Domingos, and Alon Y. Halevy. Reconciling Schemas of Disparate Data Sources: A Machine Learning Approach. In *Proceedings of the ACM SIGMOD Conference*, 2001.

[146] Anhai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Learning to map between ontologies on the semantic web. In *Proc. of the Int. WWW Conf.*, 2002.

[147] AnHai Doan, Jeffrey F. Naughton, Raghu Ramakrishnan, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, Jiansheng Huang, Warren Shen, and Ba-Quy Vuong. Information extraction challenges in managing unstructured data. *SIGMOD Record*, December 2008.

[148] Pedro Domingos and Micheal Pazzani. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 29:103–130, 1997.

[149] X. Dong, A. Y. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *Proc. of the SIGMOD Conf.*, pages 85–96, 2005.

[150] X. Dong, A. Y. Halevy, and C. Yu. Data integration with uncertainty. In *Proc. of VLDB*, 2007.

[151] Xin Dong and Alon Halevy. A Platform for Personal Information Management and Integration. In *Proc. of CIDR*, 2005.

[152] Francesco M. Donini, Maurizio Lenzerini, Daniele Nardi, and Andrea Schaerf. A hybrid system with datalog and concept languages. In E. Ardizzone, S. Gaglio,

and F. Sorbello, editors, *Trends in Artificial Intelligence*, volume LNAI 549, pages 88–97. Springer Verlag, 1991.

[153] Francesco. M. Donini, Maurizion Lenzerini, Daniele Nardi, and Werner Nutt. The complexity of concept languages. In *Proceedings of KR-91*, 1991.

[154] Mira Dontcheva, Steven M. Drucker, David Salesin, and Michael F. Cohen. Relations, cards, and search templates: user-guided web data integration and layout. In *UIST*, pages 61–70, 2007.

[155] R. B. Doorenbos, O. Etzioni, and D. S. Weld. A scalable comparison-shopping agent for the World-Wide Web. In *Agents*, 1997.

[156] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999.

[157] Oliver Duschka, Michael Genesereth, and Alon Levy. Recursive query plans for data integration. *Journal of Logic Programming, special issue on Logic Based Heterogeneous Information Systems*, 43(1):49–73, 2000.

[158] Oliver M. Duschka and Michael R. Genesereth. Answering Recursive Queries Using Views. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, 1997.

[159] Oliver M. Duschka and Michael R. Genesereth. Query planning in infomaster. In *Proceedings of the ACM Symposium on Applied Computing*, pages 109–111, San Jose, CA, 1997.

[160] Oliver M. Duschka and Alon Y. Levy. Recursive plans for information gathering. In *Proc. of the 15th Int. Joint Conf. on Artificial Intelligence(IJCAI)*, pages 778–784, 1997.

[161] Charles Elkan. A decision procedure for conjunctive query disjointness. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, Portland, Oregon, 1989.

[162] Charles Elkan. Independence of logic database queries and updates. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 154–160, 1990.

[163] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16, 2007.

[164] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.

[165] Hazem Elmeleegy, Jayant Madhavan, and Alon Y. Halevy. Harvesting relational tables from lists on the web. *PVLDB*, 2(1):1078–1089, 2009.

[166] Mohamed Y. Eltabakh, Walid G. Aref, Ahmed K. Elmagarmid, Mourad Ouzzani, and Yasin N. Silva. Supporting annotations on relations. In *EDBT*, pages 379–390, 2009.

[167] D. W. Embley, D. M. Campbell, Y. S. Jiang, S. W. Liddle, Y. Ng, D. Quass, and R. D. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data Knowl. Eng.*, 31(3):227–251, 1999.

[168] D. W. Embley, Y. S. Jiang, and Y. Ng. Record-boundary discovery in web documents. In *SIGMOD*, 1999.

[169] O. Etzioni, K. Golden, and D. Weld. Sound and efficient closed-world reasoning for planning. *Artificial Intelligence*, 89(1–2):113–148, January 1997.

[170] Ronald Fagin. Inverting schema mappings. In *Proc. of PODS*, pages 50–59, 2006.

[171] Ronald Fagin, Phokion Kolaitis, Renée J. Miller, and Lucian Popa. Data exchange: Semantics and query answering. *TCS*, 336:89–124, 2005.

[172] Ronald Fagin, Phokion G. Kolaitis, and Lucian Popa. Composing schema mappings: Second-order dependencies to the rescue. *ACM Transactions on Database Systems*, 30(4):994–1055, 2005.

[173] Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, and Wang-Chiew Tan. Quasi-inverses of schema mappings. In *Proc. of PODS*, pages 123–132, 2007.

[174] Ronald Fagin, Amnon Lotem, and Moni Naor. Optimal aggregation algorithms for middleware. *Journal of Computer and System Sciences*, 66(4):614–656, June 2003.

[175] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Society*, 64(328):1183–1210, 1969.

[176] Jonathan Finger and Neoklis Polyzotis. Robust and efficient algorithms for rank join evaluation. In *SIGMOD 2009, Proceedings of the ACM International Conference on Management of Data, June 29-July 2, 2009, Providence, Rhode Island*, pages 415–428, New York, NY, USA, 2009. ACM.

[177] Daniela Florescu, Daphne Koller, and Alon Y. Levy. Using Probabilistic Information in Data Integration. In *Proceedings of the 23rd International Conference on Very Large Databases (VLDB)*, 1997.

[178] Daniela Florescu, Alon Levy, Ioana Manolesu, and Dan Suciu. Query optimization in the presence of limited access patterns. In *Proceedings of the ACM SIGMOD Conference*, 1999.

[179] Daniela Florescu, Alon Levy, and Alberto Mendelzon. Database techniques for the world-wide web: A survey. *SIGMOD Record*, 27(3):59–74, September 1998.

[180] Daniela Florescu, Louiqa Raschid, and Patrick Valduriez. Using heterogeneous equivalences for query rewriting in multidatabase systems. In *Proceedings of the Int. Conf. on Cooperative Information Systems (COOPIS)*, 1995.

[181] J. Nathan Foster, Todd J. Green, and Val Tannen. Annotated XML: queries and provenance. In *PODS*, pages 271–280, 2008.

[182] Paul Francis, Sugih Jamin, Cheng Jin, Yixin Jin, Danny Raz, Yuval Shavitt, and Lixia Zhang. Idmaps: a global internet host distance estimation service. *IEEE/ACM Trans. Netw.*, 9(5):525–540, 2001.

[183] M. Franklin, A. Y. Halevy, and D. Maier. From databases to dataspaces: a new abstraction for information management. In *SIGMOD Record*, pages 27–33, 2005.

[184] Michael Franklin, Alon Halevy, and David Maier. From databases to dataspaces: a new abstraction for information management. *SIGMOD Rec.*, 34(4):27–33, 2005.

[185] M. Friedman and D. Weld. Efficient execution of information gathering plans. In *Proc. of the 15th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 1997.

[186] Marc Friedman, Alon Levy, and Todd Millstein. Navigational Plans for Data Integration. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 1999.

[187] Ariel Fuxman, Mauricio A. Hernández, C. T. Howard Ho, Renée J. Miller, Paolo Papotti, and Lucian Popa. Nested mappings: Schema mapping reloaded. In *VLDB*, pages 67–78, 2006.

[188] A. Gal. Why is schema matching tough and what can we do about it? *SIGMOD Record*, 35(4):2–5, 2007.

[189] A. Gal, G. Modica, H. Jamil, and A. Eyal. Automatic ontology matching using application semantics. *AI Magazine*, 26(1):21–31, 2005.

[190] Avigdor Gal. Managing uncertainty in schema matching with top-k schema mappings. *Journal of Data Semantics*, VI:90–114, 2006.

[191] Avigdor Gal, Ateret Anaby-Tavor, Alberto Trombetta, and Danilo Montesi. A framework for modeling and evaluating automatic semantic reconciliation. 2003.

[192] M. Ganesh, J. Srivastava, and T. Richardson. Mining entity-identification rules for database integration. In *Proc. of the ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 291–294, 1996.

[193] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogeneous information sources. *Journal of Intelligent Information Systems*, 8(2):117–132, March 1997.

[194] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *Database Systems: The Complete Book*. Prentice Hall, 2002.

[195] Wolfgang Gatterbauer, Magdalena Balazinska, Nodira Khoussainova, and Dan Suciu. Believe it or not: Adding belief annotations to databases. *PVLDB*, 2(1):1–12, 2009.

[196] L. Getoor and R. Miller. Data and metadata alignment, 2007. Tutorial, the Alberto Mendelzon Workshop on the Foundations of Databases and the Web.

[197] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *SOSP*, pages 29–43, 2003.

[198] C. L. Giles, K. D. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proc. of the ACM Int. Conf. on Digital Libraries*, pages 89–98, 1998.

[199] L. E. Gill. OX-LINK: The Oxford medical record linkage system. In *Proc. of the Int Record Linkage Workshop and Exposition*, 1997.

[200] François Goasdoué and Marie-Christine Rousset. Querying distributed data through distributed ontologies: A simple but scalable approach. *IEEE Intelligent Systems*, 18(5):60–65, 2003.

[201] E. M. Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.

[202] E. M. Gold. Complexity of automaton identification from given data. *Information and Control*, 37(3):302–320, 1978.

[203] Roy Goldman, Jason McHugh, and Jennifer Widom. From semistructured data to XML: Migrating the Lore data model and query language. In *ACM SIGMOD WebDB Workshop '99*, pages 25–30, 1999.

[204] Hector Gonzalez, Alon Y. Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, and Warren Shen. Google fusion tables: data management, integration and collaboration in the cloud. In *SoCC*, pages 175–180, 2010.

[205] Hector Gonzalez, Alon Y. Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen, and Jonathan Goldberg-Kidon. Google fusion tables: web-centered data management and collaboration. In *SIGMOD Conference*, pages 1061–1066, 2010.

[206] Hector Gonzalez, Alon Y. Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen, and Jonathan Goldberg-Kidon. Google fusion tables: web-centered data management and collaboration. In *SIGMOD 2010, Proceedings of the ACM International Conference on Management of Data, June 7-11, 2010, Indianapolis, Indiana*, pages 1061–1066, 2010.

[207] G. Gottlob, C. Koch, R. Baumgartner, M. Herzog, and S. Flesca. The Lixto data extraction project - back and forth between theory and practice. In *PODS*, 2004.

[208] Goetz Graefe. Query evaluation techniques for large databases. *ACM Computing Surveys*, 25(2):73–170, June 1993.

[209] L. Gravano, P. G. Ipeirotis, N. Koudas, and D. Srivastava. Text joins in an RDBMS for web data integration. In *WWW*, 2003.

[210] Luis Gravano, Panagiotis G. Ipeirotis, Nick Koudas, and Divesh Srivastava. Text joins in an RDBMS for web data integration. In *WWW*, pages 90–101, 2003.

[211] Todd J. Green. Containment of conjunctive queries on annotated relations. In *ICDT*, pages 296–309, 2009.

[212] Todd J. Green, Grigoris Karvounarakis, Zachary G. Ives, and Val Tannen. Update exchange with mappings and provenance. In *Proc. of VLDB*, 2007.

[213] Todd J. Green, Grigoris Karvounarakis, Zachary G. Ives, and Val Tannen. Update exchange with mappings and provenance. In *VLDB 2007, Proceedings of 32nd International Conference on Very Large Data Bases, September 25-27, 2007, Vienna, Austria*, 2007. Amended version available as Univ. of Pennsylvania report MS-CIS-07-26.

[214] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the Twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, June 11-13, 2007, Beijing, China*, 2007.

[215] Todd J. Green, Gerome Miklau, Makoto Onizuka, and Dan Suciu. Processing XML streams with deterministic automata and stream indexes. Available from `http://www.cs.washington.edu/homes/suciu/files/paper.ps`, February 2002.

[216] S. Grumbach and G. Mecca. In search of the lost schema. In *ICDT*, 1999.

[217] Ashish Gupta, Inderpal Singh Mumick, and V. S. Subrahmanian. Maintaining views incrementally. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993*, pages 157–166. ACM Press, 1993.

[218] Himanshu Gupta. Selection of views to materialize in a data warehouse. In *Proceedings of the International Conference on Database Theory (ICDT)*, pages 98–112, Delphi, Greece, 1997.

[219] Laura Haas, Donald Kossmann, Edward Wimmers, and Jun Yang. Optimizing queries across diverse data sources. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, Athens, Greece, 1997.

[220] Laura M. Haas, Donald Kossmann, Edward L. Wimmers, and Jun Yang. Optimizing queries across diverse data sources. In *VLDB'97, Proceedings of 23rd International Conference on Very Large Data Bases, August 25-29, 1997, Athens, Greece*, pages 276–285. Morgan Kaufman, 1997.

[221] Alon Halevy, Zachary Ives, Igor Tatarinov, and Peter Mork. Piazza: Data management infrastructure for semantic web applications. In *Proc. of the Int. WWW Conf.*, 2003.

[222] Alon Y. Halevy. Answering Queries Using Views: A Survey. *VLDB Journal*, 10(4), 2001.

[223] Alon Y. Halevy, Naveen Ashish, Dina Bitton, Michael J. Carey, Denise Draper, Jeff Pollock, Arnon Rosenthal, and Vishal Sikka. Enterprise information integration: successes, challenges and controversies. In *SIGMOD Conference*, pages 778–787, 2005.

[224] Alon Y. Halevy, Michael J. Franklin, and David Maier. Principles of dataspace systems. In *PODS*, 2006.

[225] Alon Y. Halevy, Zachary G. Ives, Jayant Madhavan, Peter Mork, Dan Suciu, and Igor Tatarinov. The piazza peer-data management system. *Transactions on Knowledge and Data Engineering, Special issue on Peer-dta management*, 2004.

[226] Alon Y. Halevy, Zachary G. Ives, Peter Mork, and Igor Tatarinov. Piazza: Data management infrastructure for semantic web applications. In *Proceedings of the Twelfth International World Wide Web Conference, Budapest, Hungary, May 20-24 2003*, pages 556–567. World-Wide Web Consortium, ACM, May 2003.

[227] Alon Y. Halevy, Zachary G. Ives, Dan Suciu, and Igor Tatarinov. Schema Mediation in Peer Data management Systems. In *Proceedings of the International Conference on Data Engineering (ICDE)*, 2003.

[228] J. Hammer, H. Garcia-Molina, S. Nestorov, R. Yerneni, M. M. Breunig, and V. Vassalos. Template-based wrappers in the tsimmis system. In *SIGMOD*, 1997.

[229] J. Hammer, J. McHugh, and H. Garcia-Molina. Semistructured data: The tsimmis experience. In *Proc. of the First East-European Symposium on Advances in Databases and Information Systems (ADBIS)*, 1997.

[230] Joachim Hammer, Hector Garcia-Molina, Svetlozar Nestorov, Ramana Yerneni, Markus M. Breunig, and Vasilis Vassalos. Template-based wrappers in the TSIMMIS system (system demonstration). In *Proceedings of the ACM SIGMOD Conference*, Tucson, Arizona, 1998.

[231] B. He and K. Chang. Statistical Schema Matching across Web Query Interfaces. In *Proceedings of the ACM SIGMOD Conference*, 2003.

[232] B. He and K. C. Chang. Statistical schema matching across web query interfaces. In *Proc. of SIGMOD*, 2003.

[233] Bin He and Kevin Chang. Automatic Complex Schema Matching across Web Query Interfaces: A Correlation Mining Approach. *TODS*, 31(1), 2006.

[234] Bin He, Kevin Chen-Chuan Chang, and Jiawei Han. Discovering complex matchings across web query interfaces: a correlation mining approach. In *KDD*, pages 148–157, 2004.

[235] Bin He, Mitesh Patel, Zeng Zhang, and Kevin Chen-Chuan Chang. Accessing the Deep Web: A survey. *Communications of the ACM*, 50(5):95–101, 2007.

[236] Hao He, Haixun Wang, Jun Yang, and Philip S. Yu. Blinks: ranked keyword searches on graphs. In *SIGMOD 2007, Proceedings of the ACM International Conference on Management of Data, June 11-14, 2007, Beijing, China*, pages 305–316, 2007.

[237] M. A. Hernandez and S. J. Stolfo. The merge/purge problem for large databases. In *Proc. of SIGMOD*, 1995.

[238] M. A. Hernndez and S. J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2:9–37, 1998.

[239] Vagelis Hristidis, Yannis Papakonstantinou, and Andrey Balmin. Keyword proximity search on XML graphs. In *ICDE*, pages 367–378, 2003.

[240] C. Hsu and M. Dung. Generating finite-state transducers for semi-structured data extraction from the web. *Inf. Syst.*, 23(8):521–538, 1998.

[241] Jiansheng Huang, Ting Chen, AnHai Doan, and Jeffrey F. Naughton. On the provenance of non-answers to queries over extracted data. *PVLDB*, 1(1):736–747, 2008.

[242] G. Huck, P. Fankhauser, K. Aberer, and E. J. Neuhold. Jedi: Extracting and synthesizing information from the Web. In *CoopIS*, 1998.

[243] R. Huebsch, B. Chun, J. Hellerstein, B. Loo, P. Maniatis, T. Roscoe, S. Shenker, I. Stoica, and A. Yumerefendi. The architecture of pier: an internet-scale query processor. In *CIDR*, pages 28–43, 2005.

[244] Ibm, inc.

[245] Ihab F. Ilyas, Walid G. Aref, and Ahmed K. Elmagarmid. Supporting top-k join queries in relational databases. In *VLDB*, pages 754–765, 2003.

[246] Ihab F. Ilyas, Walid G. Aref, Ahmed K. Elmagarmid, Hicham G. Elmongui, Rahul Shah, and Jeffrey Scott Vitter. Adaptive rank-aware query optimization in relational databases. *ACM Trans. Database Syst.*, 31(4):1257–1304, 2006.

[247] Yannis E. Ioannidis. Query optimization. *ACM Comput. Surv.*, 28(1):121–123, 1996.

[248] Yannis E. Ioannidis, Raymond T. Ng, Kyusheok Shim, and Timos K. Sellis. Parametric query optimization. *VLDB J.*, 6(2):132–151, 1997.

[249] Yannis E. Ioannidis and Raghu Ramakrishnan. Containment of conjunctive queries: Beyond relations as sets. *ACM Transactions on Database Systems*, 20(3):288–324, 1995.

[250] Panagiotis G. Ipeirotis and Luis Gravano. Distributed Search over the Hidden Web: Hierarchical Database Sampling and Selection. In *VLDB*, pages 394–405, 2002.

[251] U. Irmak and T. Suel. Interactive wrapper generation with minimal user effort. In *WWW*, 2006.

[252] Zachary Ives, Daniela Florescu, Marc Friedman, Alon Levy, and Dan Weld. An adaptive query execution engine for data integration. In *Proceedings of the ACM SIGMOD Conference*, pages 299–310, 1999.

[253] Zachary G. Ives, Todd J. Green, Grigoris Karvounarakis, Nicholas E. Taylor, Val Tannen, Partha Pratim Talukdar, Marie Jacob, and Fernando Pereira. The Orchestra collaborative data sharing system. *SIGMOD Rec.*, 2008.

[254] Zachary G. Ives, Alon Y. Halevy, and Daniel S. Weld. An XML query engine for network-bound data. *VLDB J.*, 11(4):380–402, December 2002.

[255] Zachary G. Ives, Alon Y. Halevy, and Daniel S. Weld. Adapting to source properties in processing data integration queries. In *SIGMOD 2004, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 13-18, 2004, Paris, France*, pages 395–406, June 2004.

[256] Zachary G. Ives, Craig A. Knoblock, Steven Minton, Mari Jacob, Partha Pratim Talukdar, Rattapoom Tuchindra, Jose Luis Ambite, Maria Muslea, and Cenk Gazen. Interactive data integration though smart copy & paste. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*, 2009.

[257] Zachary G. Ives and Nicholas E. Taylor. Sideways information passing for push query processing. In *Proceedings of the 24th International Conference on Data Engineering*, 2008.

[258] P. Jaccard. tude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Socit Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[259] M. A. Jaro. Unimatch: A record linkage system: User's manual. 1976. Technical Report, U.S. Bureau of the Census, Washington D.C.

[260] T. S. Jayram, Phokion Kolaitis, and Erik Vee. The containment problem for real conjunctive queries with inequalities. In *Proc. of PODS*, pages 80–89, 2006.

[261] S. Jeffery, M. Franklin, and A. Halevy. Pay-as-you-go user feedback for dataspace systems. In *Proc. of SIGMOD*, 2008.

[262] Vanja Josifovski, Marcus Fontoura, and Attila Barta. Querying XML streams. *The VLDB Journal*, 14(2):197–210, 2005.

[263] Navin Kabra and David J. DeWitt. Efficient mid-query re-optimization of sub-optimal query execution plans. In *SIGMOD 1998, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 106–117. ACM Press, 1998.

[264] Varun Kacholia, Shashank Pandit, Soumen Chakrabarti, S. Sudarshan, Rushi Desai, and Hrishikesh Karambelkar. Bidirectional expansion for keyword search on graph databases. In *VLDB 2005, Proceedings of 30th International Conference on Very Large Data Bases, August 30-September 2, 2005, Trondheim, Norway*, pages 505–516, 2005.

[265] D. V. Kalashnikov, S. Mehrotra, and Z. Chen. Exploiting relationships for domain-independent data cleaning. In *Proc. of the SDM Conf.*, 2005.

[266] Carl-Christian Kanne and Guido Moerkotte. Efficient storage of XML data. In *Proceedings of the 16th International Conference on Data Engineering, San Diego, CA USA*, page 198. IEEE Computer Society, 2000.

[267] Grigoris Karvounarakis and Zachary G. Ives. Bidirectional mappings for data and update exchange. In *WebDB*, 2008.

[268] Grigoris Karvounarakis and Zachary G. Ives. Querying data provenance. In *SIGMOD 2010, Proceedings of the ACM International Conference on Management of Data, June 7-11, 2010, Indianapolis, Indiana*, 2010.

[269] Gjergji Kasneci, Maya Ramanath, Mauro Sozio, Fabian M. Suchanek, and Gerhard Weikum. Star: Steiner-tree approximation in relationship graphs. In *ICDE*, pages 868–879, 2009.

[270] Arthur M. Keller. Algorithms for translating view updates to database updates for views involving selections, projections, and joins. In *Proceedings of the 1985 ACM SIGMOD International Conference on Management of Data, Austin, TX, May 28-31, 1985*, pages 154–163. ACM, 1985.

[271] Anastasios Kementsietsidis, Marcelo Arenas, and Rene J Miller. Mapping data in peer-to-peer systems: Semantics and algorithmic issues. In *Proc. of SIGMOD*, 2003.

[272] A. Klug. On conjunctive queries containing inequalities. *Journal of the ACM*, pages 35(1): 146–160, 1988.

[273] D. Koller and N. Friedman. *Probabilistic Graphical Models*. The MIT Press, 2009.

[274] Donald Kossmann. The state of the artin distributed query procesing. *ACM Computing Surveys*, 32(4), 2000.

[275] N. Koudas. Special issue on data quality. *IEEE Data Engineering Bulletin*, 29(2), 2006.

[276] N. Koudas, A. Marathe, and D. Srivastava. Flexible string matching against large databases in practice. In *VLDB*, 2004.

[277] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: Similarity measures and algorithms, 2006. Tutorial, the ACM SIGMOD Conference.

[278] Andries Kruger, C. Lee Giles, Frans Coetzee, Eric J. Glover, Gary William Flake, Steve Lawrence, and Christian W. Omlin. Deadliner: Building a new niche search engine. In *CIKM*, pages 272–281, 2000.

[279] N. Kushmerick. Wrapper verification. *World Wide Web*, 3(2):79–94, 2000.

[280] N. Kushmerick, R. Doorenbos, and D. Weld. Wrapper induction for information extraction. In *Proc. of the 15th Int. Joint Conf. on Artificial Intelligence(IJCAI)*, 1997.

[281] Chung T. Kwok and Daniel S. Weld. Planning to gather information. In *Proc. of the 13th National Conf. on Artificial Intelligence (AAAI)*, pages 32–39, 1996.

[282] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira. A brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93, 2002.

[283] J. D. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the Int. Conf. on Machine Learning (ICML)*, pages 282–289, 2001.

[284] Eric Lambrecht, Subbarao Kambhampati, and Senthil Gnanaprakasam. Optimizing recursive information gathering plans. In *Proc. of the 16th Int. Joint Conf on Artificial Intelligence(IJCAI)*, pages 1204–1211, 1999.

[285] T. Landers and R. Rosenberg. An overview of multibase. In *Proceedings of the Second International Symoposium on Distributed Databases*, pages 153–183. North Holland, Amsterdam, 1982.

[286] Veronique Lattes and Marie-Christine Rousset. The use of the CARIN language and algorithms for information integration: the PICSEL project. In *Proceedings of the ECAI-98 Workshop on Intelligent Information Integration*, 1998.

[287] Yoonkyong Lee, AnHai Doan, Robin Dhamankar, Alon Y. Halevy, and Pedro Domingos. imap: Discovering complex mappings between database schemas. In *Proc. of SIGMOD*, pages 383–394, 2004.

[288] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, 2002.

[289] K. Lerman, L. Getoor, S. Minton, and C. A. Knoblock. Using the structure of Web sites for automatic segmentation of tables. In *SIGMOD*, 2004.

[290] V. Levenshtein. Binay code capable of correcting deletions, insertions,and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848, 1965. original in Russian–translation in *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707-710, 1966.

[291] Alon Levy and Marie-Christine Rousset. Combining Horn rules and description logics in carin. *Artificial Intelligence*, 104:165–209, September 1998.

[292] Alon Y. Levy. Obtaining complete answers from incomplete databases. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 402–412, Bombay, India, 1996.

[293] Alon Y. Levy. Logic-based techniques in data integration. In Jack Minker, editor, *Logic-Based Artificial Intelligence*, pages 575–595. Kluwer Academic Publishers, Dordrecht, 2000.

[294] Alon Y. Levy, Alberto O. Mendelzon, Yehoshua Sagiv, and Divesh Srivastava. Answering queries using views. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 95–104, San Jose, CA, 1995.

[295] Alon Y. Levy, Anand Rajaraman, and Joann J. Ordille. Query answering algorithms for information agents. In *Proc. of the 13th National Conf. on Artificial Intelligence (AAAI)*, 1996.

[296] Alon Y. Levy, Anand Rajaraman, and Joann J. Ordille. Querying Heterogeneous Information Sources Using Source Descriptions. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, 1996.

[297] Alon Y. Levy and Yehoshua Sagiv. Queries independent of updates. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 171–181, Dublin, Ireland, 1993.

[298] Chengkai Li, Kevin Chen-Chuan Chang, Ihab F. Ilyas, and Sumin Song. RankSQL: Query algebra and optimization for relational top-k queries. In *SIGMOD 2005, Proceedings of the ACM International Conference on Management of Data, June 14-16, 2005, Baltimore, MD*, pages 131–142, 2005.

[299] X. Li, P. Morie, and D. Roth. Robust reading: Identification and tracing of ambiguous names. In *Proc. of the HLT-NAACL Conf.*, pages 17–24, 2004.

[300] X. Li, P. Morie, and D. Roth. Semantic integration in text: From ambiguous names to identifiable entities. *AI Magazine*, 26(1):45–58, 2005. A. Doan and N. Noy and A. Halevy (editors).

[301] E. P. Lim, J. Srivastava, S. Prabhakar, and J. Richardson. Entity identification in database integration. In *Proc. of the 5th Int. Conf. on Data Engineering (ICDE-93)*, pages 294–301, 1993.

[302] Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and searching web tables using entities, types and relationships. *PVLDB 3(1)*, pages 1338–1347, 2010.

[303] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-Centric Systems and Applications. Springer, 2007.

[304] B. Liu, R. L. Grossman, and Y. Zhai. Mining data records in Web pages. In *KDD*, 2003.

[305] L. Liu, C. Pu, and W. Han. XWRAP: An XML-enabled wrapper construction system for Web information sources. In *Proc. of the IEEE Intl Conf. on Data Engineering (ICDE)*, 2000.

[306] Mengmeng Liu, Svilen R. Mihaylov, Zhuowei Bao, Marie Jacob, Zachary G. Ives, Boon Thau Loo, and Sudipto Guha. SmartCIS: Integrating digital and physical environments. *ACM SIGMOD Record*, 2010.

[307] Mengmeng Liu, Nicholas E. Taylor, Wenchao Zhou, Zachary G. Ives, and Boon Thau Loo. Recursive computation of regions and connectivity in networks. In *Proceedings of the 25th International Conference on Data Engineering*, 2009.

[308] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, Matthew Jones, Edward A. Lee, Jing Tao, and Yang Zhao. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience*, pages 1039–1065, 2006.

[309] Bertram Ludäscher, Rainer Himmeröder, Georg Lausen, Wolfgang May, and Christian Schlepphorst. Managing semistructured data with *FLORID*: A deductive object-oriented perspective. *Information Systems*, 23(8), 1998. to appear.

[310] Lothar F. Mackert and Guy M. Lohman. R* optimizer validation and performance evaluation for distributed queries. In *VLDB'86, Proceedings of 12th International Conference on Very Large Data Bases, August 25-28, 1986, Kyoto, Japan*, pages 149–159. Morgan Kaufman, 1986.

[311] Lothar F. Mackert and Guy M. Lohman. R* optimizer validation and performance evaluation for local queries. In *Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 28-30, 1986*, pages 84–95. ACM Press, 1986.

[312] Jayant Madhavan, Phil Bernstein, and Erhard Rahm. Generic schema matching with cupid. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, 2001.

[313] Jayant Madhavan, Philip A. Bernstein, AnHai Doan, and Alon Y. Halevy. Corpus-based schema matching. In *Proc. of ICDE*, pages 57–68, 2005.

[314] Jayant Madhavan and Alon Halevy. Composing mappings among data sources. In *Proc. of VLDB*, 2003.

[315] Jayant Madhavan, Shawn Jeffery, Shirley Cohen, Xin Dong, David Ko, Cong Yu, and Alon Halevy. Web-scale Data Integration: You can only afford to Pay As You Go. In *CIDR*, 2007.

[316] Jayant Madhavan, David Ko, Lucja Kot, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy. Google's deep-web crawl. In *Proc. of VLDB*, pages 1241–1252, 2008.

[317] M. Magnani and D. Montesi. Uncertainty in data integration: current approaches and open problems. In *VLDB workshop on Management of Uncertain Data*, pages 18–32, 2007.

[318] M. Magnani, N. Rizopoulos, P. Brien, and D. Montesi. Schema integration based on uncertain semantic mappings. *Lecture Notes in Computer Science*, pages 31–46, 2005.

[319] Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: a system for large-scale graph processing. In *SIGMOD Conference*, pages 135–146, 2010.

[320] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[321] Amélie Marian, Nicolas Bruno, and Luis Gravano. Evaluating top-k queries over web-accessible databases. *ACM Trans. Database Syst.*, 29(2):319–362, 2004.

[322] Amélie Marian, Nicolas Bruno, and Luis Gravano. Evaluating top-k queries over web-accessible databases. *ACM Trans. Database Syst.*, 29(2):319–362, 2004.

[323] A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to noun coreference. In *Proc. of the Conf. on Advances in Neural Information Processing Systems (NIPS)*, 2004.

[324] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. A machine learning approach to building domain-specific search engines. In *IJCAI*, pages 662–667, 1999.

[325] Andrew K. McCallum, Kamal Nigam, and Lyle H. Ungar. Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. In *KDD*, 2000.

[326] R. McCann, B. K. AlShebli, Q. Le, H. Nguyen, L. Vu, and A. Doan. Mapping maintenance for data integration systems. In *VLDB*, 2005.

[327] Luke McDowell, Oren Etzioni, Alon Halevy, Henry Levy, Steven Gribble, William Pentney, Deepak Verma, and Stani Vlasseva. Enticing ordinary people onto the semantic web via instant gratification. In *Proceedings of the Second International Conference on the Semantic Web*, October 2003.

[328] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F. Moore, and Dan Suciu. The complexity of causality and responsibility for query answers and non-answers. *PVLDB*, 4(1):34–45, 2010.

[329] Alexandra Meliou, Wolfgang Gatterbauer, Suman Nath, and Dan Suciu. Tracing data errors with view-conditioned causality. In *SIGMOD 2011, Proceedings of the ACM International Conference on Management of Data, June 12-16, 2011, Athens, Greece*.

[330] Sergey Melnik, Philip A. Bernstein, Alon Y. Halevy, and Erhard Rahm. Supporting executable mappings in model management. In *Proc. of SIGMOD*, 2005.

[331] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. Similarity Flooding: A Versatile Graph Matching Algorithm. In *Proceedings of the 18th International Conference on Data Engineering (ICDE)*, 2002.

[332] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. Dremel: Interactive analysis of web-scale datasets. *PVLDB*, 3(1):330–339, 2010.

[333] Sergey Melnik, Erhard Rahm, and Phil Bernstein. Rondo: A programming platform for generic model management. In *Proc. of SIGMOD*, 2003.

[334] X. Meng, D. Hu, and C. Li. Schema-guided wrapper maintenance for Web-data extraction. In *WIDM*, 2003.

[335] Microsoft, inc.

[336] Gerome Miklau and Dan Suciu. Containment and equivalence for a fragment of XPath. *J. ACM*, 51(1):2–45, 2004.

[337] R.J. Miller, L.M. Haas, and M. Hernandez. Schema Matching as Query Discovery. In *VLDB*, 2000.

[338] Tova Milo, Serge Abiteboul, Bernd Amann, Omar Benjelloun, and Frederic Dang Ngoc. Exchanging intensional xml data. In *Proc. of SIGMOD*, pages 289–300, 2003.

[339] Tova Milo and Sagit Zohar. Using Schema Matching to Simplify Heterogeneous Data Translation. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, 1998.

[340] Hoshi Mistry, Prasan Roy, S. Sudarshan, and Krithi Ramamritham. Materialized view selection and maintenance using multi-query optimization. In *SIGMOD 2001, Proceedings of the ACM SIGMOD International Conference on Management of Data, May 21-24, 2001, Santa Barbara, California, USA*, 2001.

[341] Tom M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[342] Prasenjit Mitra, Natasha F. Noy, and Anuj R. Jaiswal. Omen: A probabilistic ontology mapping tool. In *International Semantic Web Conference*, pages 537–547, 2005.

[343] R. Mohapatra, K. Rajaraman, and S. Y. Sung. Efficient wrapper reinduction from dynamic Web sources. In *Web Intelligence*, 2004.

[344] A. E. Monge and C. Elkan. The field matching problem: Algorithms and applications. In *KDD*, 1996.

[345] A. E. Monge and C. P. Elkan. An efficient domain-independent algorithm for detecting approximatelyduplicate database records. In *Proc. of the Second ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-97)*, pages 23–29, 1997.

[346] Peter Mork, Philip A. Bernstein, and Sergey Melnik. Teaching a schema translator to produce o/r views. In *Proceedings of Entity Relationship Conference*, 2007.

[347] Amihai Motro. Integrity = validity + completeness. *ACM Transactions on Database Systems*, 14(4):480–502, December 1989.

[348] Kiran-Kumar Muniswamy-Reddy, David A. Holland, Uri Braun, and Margo I. Seltzer. Provenance-aware storage systems. In *USENIX Annual Technical Conference, General Track*, pages 43–56, 2006.

[349] I. Muslea, S. Minton, and C. A. Knoblock. A hierarchical approach to wrapper induction. In *Agents*, 1999.

[350] I. Muslea, S. Minton, and C. A. Knoblock. Hierarchical wrapper induction for semistructured information sources. *Autonomous Agents and Multi-Agent Systems*, 4(1/2):93–114, 2001.

[351] A. Nash, P. Bernstein, and S. Melnik. Composition of mappings given by embedded dependencies. *ACM Transactions on Database Systems*, 32(1), 2007.

[352] F. Naumann and M. Herschel. *An Introduction to Duplicate Detection (Synthesis Lectures on Data Management)*. Morgan & Claypool, 2010. M. Tamer Ozsu (editor).

[353] G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001.

[354] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[355] Frank Neven and Thomas Schwentick. XPath containment in the presence of disjunction, DTDs, and variables. In *Database Theory — ICDT 2003, 9th International Conference, Siena, Italy, January 8-10, 2003, Proceedings*, 2003.

[356] H. B. Newcombe, J. M. Kennedy, S. Axford, and A. James. Automatic linkage of vital records. *Science*, 130(3381):954–959, 1959.

[357] W. S. Ng, B. C. Ooi, K.-L. Tan, and A. Zhou. Peerdb: A p2p-based system for distributed data sharing. In *ICDE*, Bangalore, India, 2003.

[358] Zaiqing Nie, Ji-Rong Wen, and Wei-Ying Ma. Object-level vertical search. In *CIDR*, pages 235–246, 2007.

[359] H. Nottelmann and U. Straccia. Information retrieval and machine learning for probabilistic schema matching. *Information Processing and Management*, 43(3):552–576, 2007.

[360] Natalya F. Noy and Mark A. Musen. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2000.

[361] Natalya Freidman Noy and Mark A. Musen. Smart: Automated support for ontology merging and alignment. In *Proceedings of the Knowledge Acquisition Workshop, Banff, Canada*, 1999.

[362] Natalya Fridman Noy. Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, 33(4):65–70, 2004.

[363] Alexandros Ntoulas, Petros Zerfos, and Junghoo Cho. Downloading Textual Hidden Web Content through Keyword Queries. In *JCDL*, pages 100–109, 2005.

[364] T. Oinn, M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe. Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 2006.

[365] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. Pig Latin: a not-so-foreign language for data processing. In *SIGMOD 2008, Proceedings of the ACM International Conference on Management of Data, June 10-12, 2008, Vancouver, Canada*, pages 1099–1110, 2008.

[366] B. On, N. Koudas, D. Lee, and D. Srivastava. Group linkage. In *ICDE*, 2007.

[367] Open provenance model. http://twiki.ipaw.info/bin/view/Challenge/OPM, 2008.

[368] M. Tamer Özsu and Patrick Valduriez. *Principles of Distributed Database Systems*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 2nd edition, 1999.

[369] M. Tamer Ozsu and Patrick Valduriez. *Principles of Distributed Database Systems*. Springer, 2011.

[370] Luigi Palopoli, Domenico Sacc, G. Terracina, and Domenico Ursino. A unified graph-based framework for deriving nominal interscheme properties, type conflicts and object cluster similarities. In *Proceedings of CoopIS*, 1999.

[371] A. Parameswaran, N. Dalvi, H. Garcia-Molina, and R. Rastogi. Optimal schemes for robust web extraction. In *VLDB*, 2011.

[372] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *Proc. of the NIPS Conf.*, pages 1401–1408, 2002.

[373] Feng Peng and Sudarshan S. Chawathe. Xsq: A streaming xpath engine. *ACM Trans. Database Syst.*, 30(2):577–623, 2005.

[374] L. Philips. Hanging on the metaphone. *Computer Language Magazine*, 7(12):39–44, 1990.

[375] L. Philips. The double metaphone search algorithm. *C/C++ Users Journal*, 18(5), 2000.

[376] J. C. Pinheiro and D. X. Sun. Methods for linking and mining massive heterogeneous databases. In *Proc. of the ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 309–313, 1998.

[377] Rachel Pottinger and Philip A. Bernstein. Merging models based on given correspondences. In *Proc. of VLDB*, pages 826–873, 2003.

[378] Rachel Pottinger and Alon Halevy. Minicon: A Scalable Algorithm for Answering Queries Using Views. *VLDB Journal*, 2001.

[379] C. Pu. Key equivalence in heterogeneous databases. In *Proc. of the 1st Int. Workshop on Interoperability in Multidatabase Systems*, 1991.

[380] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.

[381] Anand Rajaraman, Yehoshua Sagiv, and Jeffrey D. Ullman. Answering queries using templates with binding patterns. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 105–112, San Jose, CA, 1995.

[382] Raghu Ramakrishnan and Johannes Gehrke. *Database Management Systems*. McGraw Hill, 2000.

[383] Vijayshankar Raman, Amol Deshpande, and Joseph M. Hellerstein. Using state modules for adaptive query processing. In *Proceedings of the 19th International Conference on Data Engineering, March 5-8, 2003, Bangalore, India*, page 353. IEEE Computer Society, IEEE Computer Society, 2003.

[384] Jun Rao, Chun Zhang, Nimrod Megiddo, and Guy M. Lohman. Automating physical database design in a parallel database. In *SIGMOD 2002, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 3-6, 2002, Madison, Wisconsin, USA*, 2002.

[385] J. Raposo, A. Pan, M. Álvarez, and J. Hidalgo. Automatically maintaining wrappers for semi-structured Web sources. *Data Knowl. Eng.*, 61(2):331–358, 2007.

[386] P. D. Ravikumar and W. Cohen. A hierarchical graphical model for record linkage. In *Proc. of the Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 454–461, 2004.

[387] Phokion Kolaitis Ronald Fagin and Lucian Popa. Data exchange: getting to the core. *ACM Transactions on Database Systems*, 30(1):174–210, 2005.

[388] Rnee Miller Ronald Fagin, Phokion Kolaitis and Lucian Popa. Data exchange: semantics and query answering. In *Proceedings of the International Conference on Database Theory (ICDT)*, pages 207–224, 2003.

[389] Mary Tork Roth, Fatma Ozcan, and Laura M. Haas. Cost models do matter: Providing cost information for diverse data sources in a federated system. In *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, Edinburgh, Scotland*, pages 599–610, 1999.

[390] Elke A. Rundensteiner, Luping Ding, Timothy M. Sutherland, Yali Zhu, Bradford Pielech, and Nishant Mehta. Cape: Continuous query engine with heterogeneous-grained adaptivity. In *VLDB*, pages 1353–1356, 2004.

[391] R. C. Russell. U.S. Patent 1,261,167.

[392] R. C. Russell. U.S. Patent 1,435,663.

[393] Y. Sagiv and M. Yannakakis. Equivalence among relational expressions with the union and difference operators. *Journal of the ACM*, 27(4):633–655, 1981.

[394] A. Sahuguet and F. Azavant. Web ecology: Recycling HTML pages as XML documents using W4F. In *WebDB (Informal Proceedings)*, 1999.

[395] Antonio Vas Salles, Jens-Peter Dittrich, Shant Krakos Karakashian, Olivier Girard, Marcos, and Lukas Blunschi. itrails: Pay-as-you-go information integration in dataspaces. In *Proc. of VLDB*, 2007.

[396] Marcos Antonio Vaz Salles, Jens-Peter Dittrich, Shant Kirakos Karakashian, Olivier René Girard, and Lukas Blunschi. iTrails: pay-as-you-go information integration in dataspaces. In *VLDB 2008, Proceedings of 33nd International Conference on Very Large Data Bases, August 26-28, 2008, Auckland, New Zealand*, pages 663–674. VLDB Endowment, 2007.

[397] S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2008.

[398] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proc. of the ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 269–278, 2002.

[399] Sunita Sarawagi and Alok Kirpal. Efficient set joins on similarity predicates. In *SIGMOD Conference*, pages 743–754, 2004.

[400] Mayssam Sayyadian, Hieu LeKhac, AnHai Doan, and Luis Gravano. Efficient keyword search across heterogeneous relational databases. In *Proceedings of the 23rd International Conference on Data Engineering*, pages 346–355, 2007.

[401] Karl Schnaitter and Neoklis Polyzotis. Evaluating rank joins with optimal cost. In *PODS*, pages 43–52, 2008.

[402] Karl Schnaitter, Joshua Spiegel, and Neoklis Polyzotis. Depth estimation for ranking query optimization. In *VLDB 2007, Proceedings of 32nd International Conference on Very Large Data Bases, September 25-27, 2007, Vienna, Austria*, pages 902–913. VLDB Endowment, 2007.

[403] Thomas Schwentick. Xpath query containment. *SIGMOD Record*, 33(1):101–109, 2004.

[404] Luc Segoufin and Victor Vianu. Validating streaming XML documents. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, June 3-5, 2002, Madison, Wisconsin USA*, pages 53–64, 2002.

[405] W. Shen, P. DeRose, R. McCann, A. Doan, and R. Ramakrishnan. Toward best-effort information extraction. In *SIGMOD*, 2008.

[406] P. Singla and P. Domingos. Object identification with attribute-mediated dependences. In *Proc. of the PKDD Conf.*, pages 297–308, 2005.

[407] John Miles Smith, Philip A. Bernstein, Umeshwar Dayal, Nathan Goodman, Terry Landers, Ken W.T. Lin, and Eugene Wong. MULTIBASE – integrating heterogeneous distributed database systems. In *Proceedings of 1981 National Computer Conference*, pages 487–499. AFIPS Press, 1981.

[408] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[409] S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272, 1999.

[410] S. Soderland, D. Fisher, J. Aseltine, and W. G. Lehnert. Crystal: Inducing a conceptual dictionary. In *IJCAI*, 1995.

[411] Michael Stonebraker. *The design and implementation of distributed INGRES*, pages 187–196. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1986. Available from `http://dl.acm.org/citation.cfm?id=4161.4170`.

[412] Michael Stonebraker, Paul M. Aoki, Witold Litwin, Avi Pfeffer, Adam Sah, Jeff Sidell, Carl Staelin, and Andrew Yu. Mariposa: A wide-area distributed database system. *VLDB J.*, 5(1):48–63, 1996.

[413] V.S. Subrahmanian, S. Adali, A. Brink, R. Emery, J. Lu, A. Rajput, T. Rogers, R. Ross, and C. Ward. HERMES: A heterogeneous reasoning and mediator system. Technical report, University of Maryland, 1995.

[414] R. L. Taft. Name search techniques. Technical report, special report No. 1, New York State Identification and Intelligence System, Albany, N.Y.

[415] Partha Pratim Talukdar, Zachary G. Ives, and Fernando Pereira. Automatically incorporating new sources in keyword search-based data integration. In *SIGMOD 2010, Proceedings of the ACM International Conference on Management of Data, June 7-11, 2010, Indianapolis, Indiana*, 2010.

[416] Partha Pratim Talukdar, Marie Jacob, Muhammad Salman Mehmood, Koby Crammer, Zachary G. Ives, Fernando Pereira, and Sudipto Guha. Learning to create data-integrating queries. In *VLDB 2008, Proceedings of 33nd International Conference on Very Large Data Bases, August 26-28, 2008, Auckland, New Zealand*, 2008.

[417] Igor Tatarinov and Alon Halevy. Efficient query reformulation in peer data management systems. In *Proc. of SIGMOD*, 2004.

[418] Igor Tatarinov, Stratis Viglas, Kevin S. Beyer, Jayavel Shanmugasundaram, Eugene J. Shekita, and Chun Zhang. Storing and querying ordered XML using a relational database system. In *SIGMOD 2002, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 3-6, 2002, Madison, Wisconsin, USA*. ACM, 2002.

[419] Nesime Tatbul, Ugur Cetintemel, Stanley B. Zdonik, Mitch Cherniack, and Michael Stonebraker. Load shedding in a data stream manager. In *VLDB 2003, Proceedings of 29th International Conference on Very Large Data Bases, September 9-12, 2003, Berlin, Germany*, pages 309–320. Morgan Kaufman, 2003.

[420] N. Taylor and Z. Ives. Reconciling while tolerating disagreement in collaborative data sharing. In *Proc. of SIGMOD*, 2006.

[421] Nicholas E. Taylor and Zachary G. Ives. Reconciling while tolerating disagreement in collaborative data sharing. In *SIGMOD 2006, Proceedings of the ACM*

International Conference on Management of Data, June 27-29, 2006, Chicago, IL. ACM, 2006.

[422] S. Tejada, C. A. Knoblock, and S. Minton. Learning object identification rules for information integration. *Inf. Syst.*, 26(8):607–633, 2001.

[423] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang 0002, Suresh Anthony, Hao Liu, and Raghotham Murthy. Hive - a petabyte scale data warehouse using hadoop. In *ICDE*, pages 996–1005, 2010.

[424] Feng Tian and David J. DeWitt. Tuple routing strategies for distributed eddies. In *VLDB 2003, Proceedings of 29th International Conference on Very Large Data Bases, September 9-12, 2003, Berlin, Germany*, pages 333–344. Morgan Kaufmann, 2003.

[425] Kai-Ming Ting and Ian H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999.

[426] Yi-Cheng Tu, Song Liu, Sunil Prabhakar, Bin Yao, and William Schroeder. Using control theory for load shedding in data stream management. In *Proceedings of the 23rd International Conference on Data Engineering*, pages 1491–1492, 2007.

[427] Rattapoom Tuchindra, Pedro Szekely, and Craig Knoblock. Building mashups by example. In *Proceedings of CHI*, pages 139–148, 2008.

[428] Jeffrey D. Ullman. *Principles of Database and Knowledge-base Systems, Volumes I, II*. Computer Science Press, Rockville MD, 1989.

[429] Jeffrey D. Ullman. Information Integration using Logical Views. In *Proceedings of the International Conference on Database Theory (ICDT)*, 1997.

[430] Tolga Urhan, Michael J. Franklin, and Laurent Amsaleg. Cost based query scrambling for initial delays. In *SIGMOD 1998, Proceedings of the ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 130–141. ACM Press, 1998.

[431] Ron van der Meyden. The complexity of querying indefinite data about linearly ordered domains. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*, pages 331–345, San Diego, CA., 1992.

[432] Panos Vassiliadis, Alkis Simitsis, and Spiros Skiadopoulos. Conceptual modeling for etl processes. In *Proceedings of the 5th ACM international workshop on*

Data Warehousing and OLAP, DOLAP '02, pages 14–21, New York, NY, USA, 2002. ACM. Available from `http://doi.acm.org/10.1145/583890.583893`.

[433] Petros Venetis, Alon Y. Halevy, Jayant Madhavan, Marius Pasca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. Recovering semantics of tables on the web. *PVLDB*, 4(9):528–538, 2011.

[434] J. Wang and F. H. Lochovsky. Data extraction and label assignment for Web databases. In *WWW*, 2003.

[435] Y. R. Wang and S. E. Madnick. The inter-database instance identification problem in integrating autonomous systems. In *Proc. of the 5th Int. Conf. on Data Engineering (ICDE-89)*, pages 46–55, 1989.

[436] M. Waterman, T. Smith, and W. Beyer. Some biological sequence metrics. *Advances in Math*, 20(4):367–387, 1976.

[437] Paul Westerman. *Data Warehousing: Using the Wal-Mart Model*. Morgan Kaufmann Publishers, 2000.

[438] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, pages 38–49, March 1992.

[439] W. E. Winkler. Improved decision rules in the Fellegi-Sunter model of record linkage, 1993. Technical Report, Statistical Research Report Series RR93/12, U.S. Bureau of the Census.

[440] W. E. Winkler. The state of record linkage and current research problems, 1999. Technical Report, Statistical Research Report Series RR99/04, U.S. Bureau of Census.

[441] W. E. Winkler. Methods for record linkage and Bayesian networks, 2002. Technical Report, Statistical Research Report Series RRS2002/05, U.S. Bureau of the Census.

[442] W. E. Winkler and Y. Thibaudeau. An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. census, 1991. Technical Report, Statistical Research Report Series RR91/09, U.S. Bureau of the Census.

[443] W. E. Winkler and Y. Thibaudeau. An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial census. 1991. Technical Report, Statistical Research Report Series RR91/09, U.S. Bureau of the Census, Washington D.C.

[444] David Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[445] Jeffrey Wong and Jason I. Hong. Making mashups with marmite: towards end-user programming for the web. In *CHI*, pages 1435–1444, 2007.

[446] Wensheng Wu, Clement Yu, AnHai Doan, and Weiyi Meng. An Interactive Clustering-based Approach to Integrating Source Query Interfaces on the Deep Web. In *SIGMOD*, 2004.

[447] Chuan Xiao, Wei Wang 0011, Xuemin Lin, and Jeffrey Xu Yu. Efficient similarity joins for near duplicate detection. In *WWW*, pages 131–140, 2008.

[448] Yahoo inc.

[449] Ling Ling Yan, Renee J. Miller, Laura M. Haas, and Ronald Fagin. Data Driven Understanding and Refinement of Schema Mappings. In *Proceedings of the ACM SIGMOD*, 2001.

[450] Beverly Yang and Hector Garcia-Molina. Improving search in peer-to-peer networks. In *ICDCS*, pages 5–14, 2002.

[451] Yuan Yu, Michael Isard, Dennis Fetterly, Mihai Budiu, Úlfar Erlingsson, Pradeep Kumar Gunda, and Jon Currey. Dryadlinq: A system for general-purpose distributed data-parallel computing using a high-level language. In *OSDI*, pages 1–14, 2008.

[452] Y. Zhai and B. Liu. Web data extraction based on partial tree alignment. In *WWW*, 2005.

[453] Y. Zhang, N. Tang, and P. A. Boncz. Efficient distribution of full-fledged XQuery. In *Engineering*, pages 565 – 576. IEEE, April 2009.