Bibliographic Notes

Dan Suciu Nilesh Dalvi

Abstract

These bibliographic notes accompany the tutorial **Foundations of Probabilistic Answers to Queries**, presented at SIGMOD'2005.

1 Part I: Applications: Managing Imprecisions

This part describes several types of imprecisions in data management.

1. Ranking query answers. The database is precise, but the query answers are imprecise.
   (a) Automated ranking of SQL results [ACDG03, CDHW04], Ontology and semantic similarity XXL [TW02].
   (b) Keyword searches: DBExplorer [CDN02], DISCOVER [HP02, HGP03], BANKS [BNH+02], XRANK [GSBS03]), Proximity Search [GSVGM98].
   (c) User preferences: Preference SQL [KK02], a logic for preferences [Cho03], fuzzy queries [FW97, Fag98], PREFER [HKP01].
   (d) Uncertain Predicates [Mot88].
2. Record Linkage. The databases are deterministic, but when one searches for matching records the results are imprecise [WWC03, Win99, GBVR03, FS69].
   (a) Data Cleaning [HS95, GFS+01, ACG02, CGGM03].
   (b) IR Style data integration: WHIRL [Coh00].
3. Quality in data Integration: Probabilistic Information [FKL97], Approximate Query Translation [CGM00], Sound/Incomplete data sources [MM01].
4. Handling Inconsistent Data. The imprecision here is due to the fact that constraints are violated. The repair model and its complexity [BCCG02, BC03, CM02], efficient query processing [FM05]).
5. Information disclosure in data sharing. Here the imprecision is the adversary's lack of knowledge of the private data. Various formalisms are in [EGS03, MS03, MS04, DMS05, DP05].

6. Other applications.
   (a) Data Lineage/Accuracy: Trio [Wid05].
   (b) Sensor Networks [CP03] and Data Acquisition [ADM05].
   (c) Personal Information Management: Haystack [KBH+03], Semex [DH05].
   (d) Information extraction from the Web (KnowItAll [ECD+04, CE05]).

2 Part II: A Probabilistic Data Semantics

Possible Worlds Semantics. The definition given in the tutorial is related to:

- Kripke semantics [Kri63] for modal logics.
- C-tables [IL84].
- Extension of First-Order logic to support probabilities with possible worlds semantics [Hal89].
- A logic for reasoning about probabilities [FH88, FHM90].
- Possible worlds semantics for database queries [Zim97].

3 Part III: Representation Formalisms

The formalisms covered in the tutorial are synthesized from the following:
Relational Models:

- Cavallo and Pittarelli [CP87]: they assume that tuples in the same relations represent disjoint events.
- Barbara et. al. [BGMP92]: generalize Cavallo and Pittarelli's model. Tuples are independent, and attributes may be inaccurate (leading to disjoint tuples). Every relation must have a set of deterministic attributes forming the key of that relation.
- Suk Lee [Lee92]: uses Dempster-Shafer theory rather than probability theory.
- Logic Programming with Interval Probabilities [NS92], assumes total ignorance while combining probabilities.
- Dey and Sarkar [DS96]: drop Barbara et al.'s key requirement, but allow only those projections that contain the key.
- The Probview system [LLRS97, RSG05]: generalize most of the previous systems. Does not assume independence, but requires user defined strategies to combine probabilities. Does not attempt to compute probabilities exactly: uses probability intervals.

- Fuhr [FR97, Fuh95]: uses events to describe possible worlds (intensional semantics).
- Conjunctive queries on probabilistic databases [DS04]: establishes a dichotomy of conjunctive queries with PTIME and #P data complexity.
- The Trio system [Wid05]: distinguishes between data lineage and its probability; further distinguishes between precision (is the tuple in the database ?) and accuracy (what is this attribute's value ?).

Non-relational models:

- ProTDB [NJ02]: XML data, assume children independent given parent.
- PXML [HGS03b]: specify complete distribution of children given parent.
- PXML with interval probabilities [HGS03a].
- Eiter et al. Object bases [ELLS01].
- Fuhr DAML [NF03].

Knowledge Bases:

- Statistics to Beliefs [BGHK96].
- Principle of Entropy/Cross Entropy for representing knowledge [GHK92, BGHK94].

4 Part IV: Foundations

1. The complexity of the probability for a boolean expression is #P-complete [Val79].
2. The data complexity for conjunctive queries over probabilistic databases is #P-complete [GGH98, DS04]. A conjunctive query has #P-hard data complexity iff it contains the subquery $P(x), Q(x, y), R(y)$: otherwise it's data complexity is PTIME [DS04].
3. 0/1 laws and Random Graphs. Glebskiĭ's et al. [GKLT69] and Fagin's [Fag76] 0/1 law for first order logic. The evolution of random graphs [ER60] following [Spe01].

Other results relevant to this part are in [SS88, Lyn92, DMS05, DS05].

5 Part V: Algorithms, Implementations

1. Monte Carlo simulation methods [KL83].
2. The Threshold Algorithm [NR99, FLN01] and it's connection to probabilistic databases..
3. Q-gram index [GIJ+01]

References

[ACDG03] S. Agrawal, S. Chaudhuri, G. Das, and A. Gionis. Automated ranking of database query results. In *CIDR*, 2003.

[ACG02] Rohit Ananthakrishna, Surajit Chaudhuri, and Venkatesh Ganti. Eliminating fuzzy duplicates in data warehouses. In *VLDB*, pages 586–596, 2002.

[ADM05] Carlos Guestrin Amol Deshpande and Samuel Madden. Using probabilistic models for data management in acquisitional environments. In *CIDR*, 2005.

[BC03] L. Bertossi and J. Chomicki. Query answering in inconsistent databases. In G. Saake J. Chomicki and R. van der Meyden, editors, *Logics for Emerging Applications of Databases*. Springer, 2003.

[BCCG02] L. Bertossi, J. Chomicki, A. Cortes, and C. Gutierrez. Consistent answers from integrated data sources. In *International Conference on Flexible Query Answering Systems*, 2002.

[BGHK94] Fahiem Bacchus, Adam Grove, Joseph Halpern, and Daphne Koller. Generating new beliefs from old. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 37–45, 1994.

[BGHK96] Fahiem Bacchus, Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. From statistical knowledge bases to degrees of belief. *Artificial Intelligence*, 87(1-2):75–143, 1996.

[BGMP92] Daniel Barbará, Hector Garcia-Molina, and Daryl Porter. The management of probabilistic data. *IEEE Trans. Knowl. Data Eng.*, 4(5):487–502, 1992.

[BNH+02] Gaurav Bhalotia, Charuta Nakhe, Arvind Hulgeri, Soumen Chakrabarti, and S. Sudarshan. Keyword searching and browsing in databases using BANKS. In *ICDE*, 2002.

[CDHW04] Surajit Chaudhuri, Gautam Das, Vagelis Hristidis, and Gerhard Weikum. Probabilistic ranking of database query results. In *VLDB*, 2004.

[CDN02] S. Chaudhuri, G. Das, and V. Narasayya. Dbexplorer: A system for keyword search over relational databases. In *Proceedings of the 18th Int. Conf. on Data Engineering, San Jose, USA*, 2002.

[CE05] Michael Cafarella and Oren Etzioni. A search engine for natural languagae applications. In *WWW*, 2005.

[CGGM03] Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti, and Rajeev Motwani. Robust and efficient fuzzy match for online data cleaning. In *SIGMOD*, pages 313–324, 2003.

[CGM00] Chen-Chuan K. Chang and Hector Garcia-Molina. Approximate query translation across heterogeneous information sources. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 566–577, 2000.

[Cho03] Jan Chomicki. Preference formulas in relational queries. *ACM Trans. Database Syst.*, 28(4):427–466, 2003.

[CM02] Jan Chomicki and Jerzy Marcinkowski. On the computational complexity of consistent query answers. In *CoRR cs.DB/0204010*, 2002.

[Coh00] William W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Trans. Inf. Syst.*, 18(3):288–321, 2000.

[CP87] Roger Cavallo and Michael Pittarelli. The theory of probabilistic databases. In *VLDB'87, Proceedings of 13th Int. Conf. on Very Large Data Bases, September 1-4, 1987, Brighton, England*, pages 71–81, 1987.

[CP03] Reynold Cheng and Sunil Prabhakar. Managing uncertainty in sensor database. *SIGMOD Rec.*, 32(4):41–46, 2003.

[DH05] Xin Dong and Alon Halevy. A platform for personal information management and integration. In *CIDR*, 2005.

[DMS05] N. Dalvi, G. Miklau, and D. Suciu. Asymptotic conditional probabilities for conjunctive queries. In *ICDT*, 2005.

[DP05] A. Deutsch and Y. Papakonstantinou. Privacy in database publishing. In *ICDT*, pages 230–245, 2005.

[DS96] Debabrata Dey and Sumit Sarkar. A probabilistic relational model and algebra. *ACM Trans. Database Syst.*, 21(3):339–369, 1996.

[DS04] Nilesh Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. In *VLDB*, 2004.

[DS05] Nilesh Dalvi and Dan Suciu. Asymptotic conditional probabilities for conjunctive queries. In *ICDT*, 2005.

[ECD+04] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in knowitall: (preliminary results). In *WWW*, pages 100–110, 2004.

[EGS03] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, 2003.

[ELLS01] Thomas Eiter, James J. Lu, Thomas Lukasiewicz, and V. S. Subrahmanian. Probabilistic object bases. *ACM Trans. Database Syst.*, 26(3):264–312, 2001.

[ER60] Paul Erdös and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kut. Int. Kozl.*, 5:17–61, 1960.

[Fag76] R. Fagin. Probabilities on finite models. *Journal of Symbolic Logic*, 41(1):50–58, 1976.

[Fag98] Ronald Fagin. Fuzzy queries in multimedia database systems. In *PODS*, pages 1–10, 1998.

[FH88] Ronald Fagin and Joseph Y. Halpern. Reasoning about knowledge and probability. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 277–293, 1988.

[FHM90] Ronald Fagin, Joseph Y. Halpern, and Nimrod Megiddo. A logic for reasoning about probabilities. *Information and Computation*, 87(1/2):78–128, 1990.

[FKL97] Daniela Florescu, Daphne Koller, and Alon Y. Levy. Using probabilistic information in data integration. In *The VLDB Journal*, pages 216–225, 1997.

[FLN01] Ronald Fagin, Amnon Lotem, and Moni Naor. Optimal aggregation algorithms for middleware. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 102–113, 2001.

[FM05] Ariel D. Fuxman and Renee J. Miller. First-order query rewriting for inconsistent databases. In *ICDT*, pages 337–351, 2005.

[FR97] Norbert Fuhr and Thomas Roelleke. A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Trans. Inf. Syst.*, 15(1):32–66, 1997.

[FS69] I. P. Fellegi and A. B. Sunter. A theory for record linkage. In *Journal of the American Statistical Society*, volume 64, pages 1183–1210, 1969.

[Fuh95] Norbert Fuhr. Probabilistic datalogȧ logic for powerful retrieval methods. In *SIGIR*, pages 282–290, 1995.

[FW97] Ronald Fagin and Edward L. Wimmers. Incorporating user preferences in multimedia queries. In *ICDT*, pages 247–261, 1997.

[GBVR03] Lifang Gu, Rohan Baxter, Deanne Vickers, and Chris Rainsford. Record linkage: Current practice and future directions. In *CMIS Technical Report No. 03/83*, 2003.

[GFS+01] Helena Galhardas, Daniela Florescu, Dennis Shasha, Eric Simon, and Cristian-Augustin Saita. Declarative data cleaning: Language, model, and algorithms. In *VLDB*, pages 371–380, 2001.

[GGH98] Erich Gradel, Yuri Gurevich, and Colin Hirch. The complexity of query reliability. In *Symposium on Principles of Database Systems*, pages 227–234, 1998.

[GHK92] Adam J. Grove, Joseph Y. Halpern, and Daphne Koller. Random worlds and maximum entropy. In *Logic in Computer Science*, pages 22–33, 1992.

[GIJ+01] L. Gravano, P. Ipeirotis, H.V. Jagadish, N. Koudas, S. Muthukrishnan, and D. Srivastava. Approximate string joins in a database (almost) for free. In *VLDB*, 2001.

[GKLT69] Y. V. Glebskiĭ, D. I. Kogan, M. I. Liogon'kiĭ, and V. A. Talanov. Range and degree of realizability of formulas in the restricted predicate calculus. *Kibernetika*, 2:17–28, 1969. [Engl. Transl. Cybernetics, vol. 5, 142–154 (1972)].

[GSBS03] Lin Guo, Feng Shao, Chavdar Botev, and Jayavel Shanmugasundaram. Xrank: Ranked keyword search over xml documents. In *Proceedings of the 2003 ACM SIGMOD Int. Conf. on Management of Data, San Diego, California, USA, June 9-12, 2003*, pages 16–27, 2003.

[GSVGM98] Roy Goldman, Narayanan Shivakumar, Suresh Venkatasubramanian, and Hector Garcia-Molina. Proximity search in databases. In *VLDB*, pages 26–37, 1998.

[Hal89] Joseph Y. Halpern. An analysis of first-order logics of probability. In *IJCAI*, pages 1375–1381, Detroit, US, 1989.

[HGP03] Vagelis Hristidis, Luis Gravano, and Yannis Papakonstantinou. Efficient ir-style keyword search over relational databases. In *VLDB*, pages 850–861, 2003.

[HGS03a] Edward Hung, Lise Getoor, and V. S. Subrahmanian. Probabilistic interval xml. In *ICDE*, 2003.

[HGS03b] Edward Hung, Lise Getoor, and V. S. Subrahmanian. Pxml: A probabilistic semistructured data model and algebra. In *ICDE*, 2003.

[HKP01] Vagelis Hristidis, Nick Koudas, and Yannis Papakonstantinou. Prefer: a system for the efficient execution of multi-parametric ranked queries. In *SIGMOD*, pages 259–270, 2001.

[HP02] V. Hristidis and Y. Papakonstantinou. Discover: Keyword search in relational databases. In *Proc. 28th Int. Conf. Very Large Data Bases, VLDB*, 2002.

[HS95] Mauricio Hernandez and Salvatore Stolfo. The merge/purge problem for large databases. In *SIGMOD*, pages 127–138, 1995.

[IL84] T. Imielinski and W. Lipski. Incomplete information in relational databases. *Journal of the ACM*, 31:761–791, October 1984.

[KBH+03] David Karger, Karun Bakshi, David Huynh, Dennis Quan, and Vineet Sinha. Haystack: A customizable general-purpose information. management tool for end users of semistructured. In *CIDR*, 2003.

[KK02] Werner Kiebling and Gerhard Kostler. Preference sql - design, implementation, experiences. In *VLDB*, pages 990–1001, 2002.

[KL83] Richard Karp and Michael Luby. Monte-carlo algorithms for enumeration and reliability problems. In *Proceedings of the annual ACM symposium on Theory of computing*, 1983.

[Kri63] S. A. Kripke. Semantic analysis of modal logic. i: Normal propositional calculi. In *Zeitschrift fiir Mathematische Logik und Grundlagen der Mathematik, 9:67-96*, 1963.

[Lee92] Suk Kyoon Lee. An extended relational database model for uncertain and imprecise information. In *VLDB*, pages 211–220, 1992.

[LLRS97] Laks V. S. Lakshmanan, Nicola Leone, Robert Ross, and V. S. Subrahmanian. Probview: a flexible probabilistic database system. *ACM Trans. Database Syst.*, 22(3):419–469, 1997.

[Lyn92] James F. Lynch. Probabilities of sentences about very sparse random graphs. *Random Struct. Algorithms*, 3(1):33–54, 1992.

[MM01] Alberto O. Mendelzon and George A. Mihaila. Querying partially sound and complete data sources. In *PODS*, pages 162–170, 2001.

[Mot88] Amihai Motro. Vague: a user interface to relational databases that permits vague queries. *ACM Trans. Inf. Syst.*, 6(3):187–214, 1988.

[MS03] Gerome Miklau and Dan Suciu. Controlling access to published data using cryptography. In *VLDB*, pages 898–909, 2003.

[MS04] Gerome Miklau and Dan Suciu. A formal analysis of information disclosure in data exchange. In *SIGMOD*, 2004.

[NF03] H. Nottelmann and N. Fuhr. Combining DAML+OIL, XSLT and probabilistic logics for uncertain schema mappings in MIND. In *ECDL*, 2003.

[NJ02] Andrew Nierman and H. V. Jagadish. ProTDB: Probabilistic data in XML. In *VLDB*, 2002.

[NR99] S. Nepal and M. V. Ramakrishna. Query processing issues in image (multimedia) databases. In *ICDE*, pages 22–29, 1999.

[NS92] Raymond T. Ng and V. S. Subrahmanian. Probabilistic logic programming. *Information and Computation*, 101(2):150–201, 1992.

[RSG05] R. Ross, V.S. Subrahmanian, and J. Grant. Aggregate operators in probabilistic databases. *Journal of the ACM*, 52(1):54–101, 2005.

[Spe01] Joel Spencer. *The Strange Logic of Random Graphs*. Springer, 2001.

[SS88] J. Spencer and S. Shelah. Zero-one laws for sparse random graphs. *J. Amer. Math. Soc.*, pages 97–115, 1988.

[TW02] Anja Theobald and Gerhard Weikum. The xxl search engine: ranked retrieval of xml data using indexes and ontologies. In *SIGMOD*, pages 615–615, 2002.

[Val79] L. Valiant. The complexity of enumeration and reliability problems. *SIAM J. Comput.*, 8:410–421, 1979.

[Wid05] Jennifer Widom. Trio: A system for integrated management of data, accuracy, and lineage. In *CIDR*, 2005.

[Win99] William Winkler. The state of record linkage and current research problems. In *Technical Report, Statistical Research Division, U.S. Bureau of the Census*, 1999.

[WWC03] Stephen E. Fienberg William W. Cohen, Pradeep Ravikumar. A comparison of string distance metrics for name-matching tasks. In *IIWeb*, pages 73–78, 2003.

[Zim97] E. Zimanyi. Query evaluation in probabilistic databases. *Theoretical Computer Science*, 171(1-2):179–219, 1997.