Ph.D. Topic 2014

Learning Approaches in Dynamic Data Management

Supervisor

Prof. Pierre Senellart, Télécom ParisTech pierre.senellart@telecom-paristech.fr

Co-supervisor(s)

Prof. Stéphane Bressan, NUS
steph@nus.edu.sg

Image & Pervasive Access Lab
1 Fusionopolis Way
#21-01 Connexis, South Tower
Singapore 138632

Tel. (65) 6408 2542
Director. (65) 6408 2536
Fax. (65) 6776 1378

secretariat@ipal.cnrs.fr
www.ipal.cnrs.fr

Background and Topic Description

Dynamic Data Management

A number of data management tasks are by nature *dynamic*: in order to achieve a *goal* (e.g., answer a query or a query load) effectively and efficiently, *actions* are carried out that make the database evolve (e.g., add new information, enrich existing information, change the physical organization of the data). Actions typically have a *cost* (computational cost, network access, I/O, actual monetary cost, etc.). The main issue of dynamic data management can be formulated as follows: *given my current database, what is the next best action to carry out to reach my goal?* Here, *best* is to be formulated in terms of the cost of the action, and effectiveness and efficiency of reaching the goal once the action is performed. The following examples all follow this model:

Self-tuning databases [1]
: Given a workload of queries, not known in advance, modify the physical organization of the database to efficiently perform query answering, while minimizing the cost of the reorganization.

Data cleaning [2]
: Given a business decision to be made, clean and curate a database by performing costly requests to human experts so as to effectively reach a decision, while minimizing the cost of the accesses.

Enrichment by Web services [3]
: Given a query over an incomplete knowledge base, call Web services to enrich the knowledge base so as to answer the query, while minimizing the number of Web requests made.

Crowd mining [4]
: Given a mining intent, ask questions on crowdsourcing platforms so as to effectively and efficiently answer this intent, while minimizing the budget spent in querying the crowd.

Focused crawling [5]
: Given a topic, progressively obtain Web pages relevant to this topic, while minimizing the number of HTTP requests made.

Reinforcement learning

Determining *the next best thing to do* under an evolving knowledge of the real world is the focus of *reinforcement learning* [6]. Its objective is to maximize the cumulative reward obtained when performing some actions, each action leading to an individual reward and to a new state, usually in a stochastic manner. *Markov decision processes* [7], a.k.a. MDPs, are a common model for reinforcement learning scenarios, where each action leads to a new state and a given reward according to a probability distribution that must be learned. This implies an inherent tradeoff between exploration (trying out new actions leading to new states and to potentially high rewards) and exploitation (performing actions already known to yield high rewards), a compromise in, e.g., the stateless model of multi-armed bandits [8]. The use of MDPs for modeling data cleaning tasks has been raised in [2]; as discussed by the authors, there is no straightforward way to do that because of the huge state space. There are, however, many challenges to using MDPs for data management: (i) the state space is typically huge, representing all possible partial knowledge on the data; (ii) states have complex structures, namely that of the data; (iii) rewards are typically delayed, as queries may only be answerable after a long sequence of actions.

Image & Pervasive Access Lab
1 Fusionopolis Way
#21-01 Connexis, South Tower
Singapore 138632

Tel. (65) 6408 2542
Director. (65) 6408 2536
Fax. (65) 6776 1378

secretariat@ipal.cnrs.fr
www.ipal.cnrs.fr

Active learning

Active learning [9] can be seen as a subfield of machine learning concerned with how to optimally use a (costly) oracle, in an interactive manner, to label training data that will be used to build a learning model, e.g., a classifier. As in dynamic data management, the system has to optimize the number of accesses made to the oracle in order to gain certainty on the output of a learning task, with a trade-off between the cost to call the oracle, and the cost of errors on the task. However, in most of the active learning literature, the cost model is very basic (uniform or fixed-value costs), though some works [10 ,11] consider more realistic costs. Also, oracles are usually assumed to be perfect with only a few exceptions [11]. Perhaps a more discriminating feature of dynamic data management with respect to the active learning literature is that data is discovered as it is accessed: there is no fixed set of data points with known features to choose from, but there are structural constraints on the data. Besides, data cannot be accessed arbitrarily but must follow source access constraints. Finally, in our setting there is an ultimate, potentially complex, goal in mind, such as a logical query with actions revealing just a part of the knowledge. This contrasts with the classical active learning scenario where the oracle returns a labeled example, and labeling is the only goal of the system.

Expected deliverables

The goal of this PhD research will be to explore the applicability of approaches from reinforcement learning and active learning to dynamic data management and to propose solutions to the challenges inherent in dynamic data management (partial knowledge of the structure of the data, huge number of states, complex logical structures of states and complex goal, delayed rewards, complex cost models). The focus will be on foundational, basic research, with the goal of exhibiting a framework for generic dynamic data management tasks. This framework may be instantiated to several dynamic data management applications in the course of the PhD research.

Keywords

Data Management, Dynamic Data, Querying, Reinforcement learning, Active learning

Image & Pervasive Access Lab
1 Fusionopolis Way
#21-01 Connexis, South Tower
Singapore 138632

Tel. (65) 6408 2542
Director. (65) 6408 2536
Fax. (65) 6776 1378

secretariat@ipal.cnrs.fr
www.ipal.cnrs.fr

Environment

The 3-year PhD thesis will be carried out within the French–Singaporean lab IPAL[1] and jointly supervised by Pierre Senellart, Professor at Télécom ParisTech, and Stéphane Bressan, Associate Professor at the National University of Singapore. The PhD candidate will be hosted at NUS in Singapore and interact with other members of the IPAL laboratory, of the Computer Science department at NUS, and of the DBWeb[2] team at Télécom ParisTech. The PhD candidate will be registered at the EDITE graduate school[3] in Paris and, upon successful completion of his research, be awarded a PhD diploma from Télécom ParisTech.

Applicant profile

- Master's Degree or Engineer Student (last year of studies), specialization in computer science or computer engineering
- Research experience in the area of data management, machine learning, or data mining
- Strong motivation towards this challenging project.
- Open to work with both French and Singaporean scientists
- Availability for starting at the Fall of 2014 (flexible)

Gratification: Compliant to French Regulation on Ph.D. students (Contrat doctoral)

1 http://ipal.i2r.a-_star.edu.sg/
2 http://dbweb.enst.fr/
3 https://edite-de-paris.fr/spip/

References

[1] Surajit Chaudhuri and Vivek R. Narasayya. Self-tuning database systems: A decade of progress. In VLDB, pages 3–14, 2007.

[2] Michael Benedikt, Philip Bohannon, and Glenn Bruns. Data cleaning for decision support. In CleanDB, 2006.

[3] Nicoleta Preda, Gjergji Kasneci, Fabian M. Suchanek, Thomas Neumann, Wenjun Yuan, and Gerhard Weikum. Active knowledge: dynamically enriching rdf knowledge bases by web services. In SIGMOD Conference, pages 399–410, 2010.

[4] Yael Amsterdamer, Yael Grossman, Tova Milo, and Pierre Senellart. Crowd mining. In Proc. SIGMOD, pages 241–252, New York, USA, June 2013.

[5] Georges Gouriten, Silviu Maniu, and Pierre Senellart. Scalable, generic, and adaptive systems for focused crawling. In Proc. Hypertext, Santiago, Chile, September 2014.

[6] Richard S. Sutton and Andrew G. Barto. Reinforcement Learning. MIT Press, 1998.

[7] Martin L. Puteman. Markov Decision Processes. Wiley, 2005.

[8] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theor. Comput. Sci., 410(19), 2009.

[9] Burr Settles. Active Learning. Morgan & Claypool Publishers, 2012.

[10] B. Settles, M. Craven, and L. Friedland. Active learning with real annotation costs. In Cost-Sensitive Learning, 2008.

[11] Pinar Donmez and Jaime G. Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In CIKM, 2008.

Image & Pervasive Access Lab
1 Fusionopolis Way
#21-01 Connexis, South Tower
Singapore 138632

Tel. (65) 6408 2542
Director. (65) 6408 2536
Fax. (65) 6776 1378

secretariat@ipal.cnrs.fr
www.ipal.cnrs.fr