August 25 and 26 • Rio de Janeiro - Brazil

33rd Brazilian Symposium on DATABASES

Proceedings

Realização: SBC Sociedade Brasileira de Computação
Organização: IME
Apoio: CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico; CAPES
Patrocínio: Google
Apoio Institucional: PUC RIO

XXXIII BRAZILIAN SYMPOSIUM ON DATABASES (SBBD 2018)

August 25-26, 2018
Rio de Janeiro, RJ, Brazil

PROCEEDINGS Sociedade Brasileira de Computação (SBC)

SUPPORT
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

SPONSOR
Google

ACADEMIC SUPPORT
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

005.74 Brazilian Symposium on Databases (SBBD 2018) (25.: 2018: Rio de Janeiro, RJ)

B827 Proceedings of 33nd Brazilan Symposium on Databases (SBBD 2018): August 25-26, 2018 – Rio de Janeiro, RJ, Brazil; organizadores: Bernadette Farias Lóscio; Carina Friedrich Dorneles; Maria Camila Nardini Barioni – Rio de Janeiro: SBC, 2018.

ISSN: 2016-5170
volume 01
316p.
Modo de acesso: http://sbbd.org.br/2018/

1. Computação - Congressos. 2. Bases de Dados – Congressos. l. Lóscio, Bernadette Farias. II. Dorneles, Carina Fiedrich. III. Barioni, Maria Camila Nardini. IV. Sociedade Brasileira de Computação. V. Título.

Message from the Local Organization Committee Chair

Welcome to the 33rd Brazilian Symposium on Databases! The Brazilian Symposium on Databases is the official database event of the Brazilian Computer Society (SBC) and the largest venue in Latin America for presentation and discussion of research results in the database domain. The 33rd edition of the symposium (SBBD 2018) was held in Rio de Janeiro, RJ, from August 25th to 26th, 2018. The local organization was performed by Instituto Militar de Engenharia (IME).

This year, for the first time, SBBD was held right before and at the same location of the 44thInternational Conference on Very Large Databases Conference (VLDB 2018), one of the main international conferences in the database area. This was a great opportunity for gathering national and international database researchers.

The SBBD 2018 program offered a variety of activities, suited for an audience ranging from undergraduate to Ph.D. students, database professionals, practitioners and researchers. The Program included: 4 invited talks and 2 tutorials, presented by distinguished speakers from Brazil, USA and Switzerland; 6 technical sessions; demos and applications session; posters sessions; thesis and dissertations workshop.

The excellence of SBBD 2018 program is the result of the competence and effort of a large community, which we gratefully acknowledge. The various sections of these proceedings list in detail those that contributed to the SBBD 2018 edition. We thank the symposium chairs and our colleagues of the local organization committee who donated their precious time to make SBBD 2018 a reality. We specially thank IME and its Post-graduation Programs (PGSC and PGED), which allowed their staff and students to help on the many tasks of the event preparation. We are also grateful to the SBC board for their support and to the steering committee members for their help, advice and support. Further, we thank the program committee members and external reviewers for the high quality reviews, and the authors who submitted their papers to SBBD 2018. Finally, we are grateful to our sponsors. Without their support we would not be able to organize this annual event that brings together our community.

We had a great time hosting SBBD 2018 in Rio de Janeiro!

Maria Claudia Reis Cavalcanti (IME)
Local Organization Chair – SBBD 2018

Foreword

It is a great pleasure to introduce the Proceedings of the Brazilian Symposium on Databases (SBBD) with the full, short and industrial papers accepted for presentation at the 33rd edition of the symposium. SBBD 2018 was held in Rio de Janeiro, in the state of Rio de Janeiro, Brazil, from August 25th to August 26th, 2018. It was organized by Seção de Engenharia de Computação, Instituto Militar de Engenharia (IME). This edition was held in two days of August instead of the traditional four days of October to motivate and to facilitate the participation of the database Brazilian community in the VLDB 2018 (44th International Conference on Very Large Data Bases). VLDB 2018 also took place in Rio de Janeiro, Brazil, from August 27th to August 31st.

SBBD is the official database event of the Brazilian Computer Society (SBC) and the largest venue in Latin America for presentation and discussion of research results in the databases domain. Along with technical sessions, SBBD includes invited talks and tutorials given by distinguished speakers from the national and international research communities. SBBD regularly promotes a demos and applications session, and a thesis and dissertations workshop as co-located events.

All papers presented in technical sessions during the event reported interesting results or proposed novel thought-provoking ideas in several subjects on the databases and related areas. For the 2018 edition, SBBD accepted six categories of submissions: JIDM articles, full papers, short papers, industrial papers, vision papers, and distinguished published papers.

Submissions to the JIDM category could be made throughout the year. The review process was conducted by the editorial board of JIDM, leaded by the editor-in-chief Angelo Brayner. This year, no JIDM articles were presented during the event. The Full papers track had one cycle of submissions with deadline in March. The review process for the SBBD full papers was performed in one round with a rebuttal phase. Authors were initially notified with the reviews and had a few days for answering the reviewers' comments during the rebuttal phase. After evaluating the rebuttal comments during the discussion period, a final decision was achieved. Out of 40 submitted papers and 17 were accepted as full papers(acceptance rate of 42,5%).

Short, Industrial and Vision papers track had a single cycle of submissions with deadline in June. There was no rebuttal phase. This year's Short, Industrial and Vision track features 15 short papers and 2 industrial papers. Initially, there were 36 submissions in the track, 29 in the Short category and 7 in the Industrial category. The best full and short papers received award certificates during the event and will be invited to submit extended versions of their papers to JIDM.

Distinguished Published Papers is a category of submission introduced in SBBD 2017. It aimed to attract the best papers of the Brazilian community, published or accepted for publication by a first-class database conference, and give the authors the opportunity to present their work during the event. There was one submission that was accepted by the SBBD Steering Committee, on the grounds of relevance of the original publication venue to the database community.

The topics with more submissions among full and short papers (according to the author's selection from the Topics of Interest) were: Data Analytics and Data Visualization (17 submissions), Data on the Web (16 submissions), Algorithms and Techniques for Data Mining (14 submissions), Semantic Web, Linked Data and Ontologies (14 submissions), Data Modeling (11 submissions), Information Integration and Interoperability (9 submissions), Social Data Processing (7 submissions), Graph Databases (7 submissions) and Authorization, Privacy and Security in Databases (7 submissions).

The Proceedings of SBBD are the result of the collective effort of a large community, which we gratefully acknowledge. We thank the SBBD 2018 local organization committee and its symposium chairs, who worked hard to guarantee an outstanding event. We do not have enough words to thank all committee members and external reviewers for their commitment and high quality reviews. We are also grateful to the steering committee members for their help, advice and support. Finally, we are grateful to the authors who submitted their work to SBBD 2018.

Bernadette F. Lóscio (UFPE)
Program Chair – SBBD 2018 – Full Papers

Carina F. Dorneles (UFSC)
Program Chair – SBBD 2018 – Short, Vision and Industrial Papers

Editorial for the Tutorials Track

Tutorials at the Brazilian Symposium on Databases (SBBD) have the goal to present introductory and advanced discussions on topics within the area of databases. Introductory tutorials target an audience consisting of advanced undergraduate and graduate students, as well as attendees from industry. Advanced tutorials, on the other hand, cover a state-of-the-art topic, motivating and exposing potential research paths.

The two accepted tutorials this year are related to relevant nowadays topics. The first one is entitled "In-Memory Analytic DBMSs: Design and Lessons Learned" and will discuss how in-memory analytic DBMSs are designed and built and outline the architecture of some state-of-art in-memory database systems, stressing the characteristics that differentiate them from the traditional DBMS design literature. It will be presented by Pedro Eugenio Rocha Pedreira, Software Engineer at Facebook focused on database research.

The second tutorial, "Coleta, Integração e Pré-processamento de Dados de Múltiplas Fontes", will be presented by Natércia Aguilar (MSc. student), Michele A. Brandão (Postdoctoral researcher) and Mirella M. Moro (Associate Professor) from Federal University of Minas Gerais (UFMG). They will talk about the main challenges related to collecting, integrating and preprocessing data from multiple Web sources. This tutorial will address these three issues in an integrated way with a focus on practical and research questions.

This year we had four excellent submissions, and I would like to thank all the authors of submitted proposals. I am also grateful to the tutorials program committee members for the high-quality reviews. Also, I would like to invite all of you to attend and take advantage of the selected tutorials.

Maria Camila Nardini Barioni (UFU)
Program Chair – SBBD 2018 – Tutorials

Full Papers – Technical Committee

Program Chair

Bernadette F. Lóscio (UFPE)

Program Committee

Alexandre Plastino (UFF)
Altigran Soares da Silva (UFAM)
Ana Carolina Salgado (UFPE)
André Santanchè (UNICAMP)
Angelo Brayner (UFC)
Caetano Traina Jr (ICMC/USP)
Carina F. Dorneles (UFSC)
Carmem Hara (UFPR)
Celso Hirata (ITA)
Clodoveu Davis (UFMG)
Cristina Ciferri (USP)
Damires Souza (IFPB)
Daniel de Oliveira (UFF)
Daniel Kaster (UEL)
Denio Duarte (UFFS)
Divesh Srivastava (AT&T Labs-Research, EUA)
Duncan Ruiz (PUCRS)
Edleno Moura (UFAM)
Eduardo Ogasawara (CEFET/RJ)
Elaine Sousa (USP)
Fernanda Baião (UNIRIO)
Genoveva Vargas-Solar (CNRS, França)
Humberto Razente (UFU)
Javam Machado (UFC)
João Eduardo Ferreira (IME/USP)
Jonice de Oliveira Sampaio Oliveira (IM/UFRJ)
José Palazzo Moreira de Oliveira (UFRGS)
Karin Becker (UFRGS)
Kelly Braghetto (IME/USP)
Khalid Belhajjame (U. Paris Dauphine, França)
Luciano Barbosa (UFPE)
Marco Antonio Casanova (PUC-Rio)
Marcos Gonçalves (UFMG)
Maria Camila Nardini Barioni (UFU)
Mirella Moro (UFMG)
Mirian Halfeld-Ferrari (U. d'Orleans, França)
Pedro Eugenio Rocha Pedreira (Facebook Inc.)
Renata Galante (UFRGS)
Renato Fileto (UFSC)
Ricardo Ciferri (UFSCar)
Ricardo Torres (UNICAMP)
Ronaldo Mello (UFSC)
Sergio Lifschitz (PUC-Rio)
Valéria C. Times (UFPE)
Vanessa Braganholo (UFF)
Vania Bogorny (UFSC)
Vaninha Vieira (UFBA)
Wagner Meira Jr. (UFMG)
Zoubida Kedad (UVSQ, França)

Additional Reviewers

Anderson Chaves Carniel (USP)
Christian Quevedo (PUCRS)
Diego Pessoa (UFPE)
Jonnathan Carvalho (IFFluminense)
Levy Souza (UFMG)
Marcelo Iury S. Oliveira (UFRPE)
Mauro Roisenberg (UFSC)
Nielsen Machado (PUCRS)
Pablo Silva (UFF)
Pedro Losco Takecian (IME/USP)
Priscilla Vieira (UFPB)

Short, Vision and Industrial Papers – Technical Committee

Program Chair

Carina F. Dorneles (UFSC)

Program Committee

Alessandreia Oliveira (UFJF)
Altigran Soares da Silva (UFAM)
Ana Carolina Almeida (UERJ)
Anderson Ferreira (UFOP)
Angelo Brayner (UFC)
Bernadette F. Lóscio (UFPE)
Carlos Eduardo Pires (UFCG)
Carmem Hara (UFPR)
Cristina Ciferri (USP)
Damires Souza (IFPB)
Daniel de Oliveira (UFF)
Daniel Kaster (UEL)
Daniela Barreiro Claro (UFBA)
Deise Saccol (UFSM)
Denio Duarte (UFFS)
Duncan Ruiz (PUCRS)
Eduardo Borges (FURG)
Eduardo de Almeida (UFPR)
Eduardo Ogasawara (CEFET/RJ)
Elaine Sousa (USP)
Eveline Sacramento (FUNCEME)
Fernanda Baião (UNIRIO)
Flávio R. C. Sousa (UFC)
Helena Ribeiro (UCS)
Humberto Razente (UFU)
João Eduardo Ferreira (IME/USP)
Jonas Dias (DELL EMC)
Jonice de Oliveira Sampaio Oliveira (IM/UFRJ)
José Antonio Macêdo (UFC)
José de Aguiar Moraes Filho (UNIFOR)
José Monteiro (UFC)
José Palazzo Moreira de Oliveira (UFRGS)
Karin Becker (UFRGS)
Kelly Braghetto (IME/USP)
Luciano Barbosa (UFPE)
Luiz Celso Gomes Jr (UTFPR)
Luiz Manoel Rocha Gadelha Júnior (LNCC)
Maria Camila Nardini Barioni (UFU)
Maristela Holanda (UnB)
Michele Brandão (UFMG)
Mirella Moro (UFMG)
Moisés Carvalho (UFAM)
Pedro Eugenio Rocha Pedreira (Facebook Inc.)
Raquel Stasiu (PUCPR / UTFPR)
Raqueline Penteado (UEM)
Rebeca Schroeder (UDESC)
Renata Galante (UFRGS)
Renato Fileto (UFSC)
Robson Cordeiro (USP)
Robson Fidalgo (UFPE)
Sergio Lifschitz (PUC-Rio)
Sergio Mergen (UFSM)
Thiago Henrique Silva (UTFPR)
Ticiana Coelho da Silva (UFC)
Valéria C. Times (UFPE)
Vanessa Braganholo (UFF)
Vaninha Vieira (UFBA)

Additional Reviewers

Marcelo Iury S. Oliveira (UFRPE)
Eduardo Pena (UFPR)
Guilherme Queiroz Vasconcelos (USP)
Nielsen Luiz Rechia Machado (PUCRS)

Tutorials – Technical Committee

Program Chair

Maria Camila Nardini Barioni (UFU)

Program Committee

Agma Juci Machado Traina (ICMC/USP)
Ana Carolina Salgado (UFPE)
Caetano Traina Jr (ICMC/USP)
Javam Machado (UFC)
Marta Mattoso (COPPE/UFRJ)

Table of Contents

Full Papers

Performance Evaluation and Benchmarking and; Concurrency Control and Recovery

Short, Vision and Industrial Papers

Tutorials

Keynotes

Author Index 315

SBBD 2018

Tutorials

In-Memory Analytic DBMSs: Design and Lessons Learned

Pedro Eugenio Rocha Pedreira

Facebook Inc.
1 Hacker Way
Menlo Park, CA, USA
pedroerp@fb.com

Abstract

In-memory DBMSs are systems that primarily rely on main memory for data storage. Many open source and commercial in-memory OLTP DBMSs were developed and studied in the last decade, such as VoltDB, Oracle TimesTen and Hekaton, as the amount of main memory available outpaced the size of most transactional working sets. As memory density kept increasing, a more recent trend is to leverage in-memory systems to speedup critical analytical workloads. In this scenario, in-memory analytic DBMSs have become a viable option to speedup analytic queries to important datasets.

Today, there are three main configurations in which in-memory analytic DBMSs can be used. First, these systems can be seen as *accelerators* for critical data marts which are part of much larger – and slower – data warehouses. These use cases focus on fast reporting and allowing users to interactively slice and dice portions of the data warehouse data. Second, these systems can be used to incrementally consume logs from messaging systems such as Kafka, providing analytics over realtime data and minimizing data latency. Third, some systems like MemSQL, SAP Hana and Hyper take a different set of trade-offs and are able to handle both OLTP and OLAP workloads in a hybrid configuration, also known as HTAP (Hybrid Transactional/Analytic Processing).

Besides the differences from traditional disk based analytic systems, in-memory analytic DBMSs also differ from in-memory OLTP DBMSs in many aspects. Since analytic queries are mostly composed of large memory scans, the lack of data locality, instruction and data cache misses, virtual and non-inlined function calls, remote NUMA accesses and branch mispredictions all have a non-trivial impact on query performance. JIT compilation at query-time, therefore, becomes a first class citizen to tighten scan loops, as opposed to the traditional approach based on stored procedures. In addition, the traditional *volcano* iterator chaining model, which is usually implemented using virtual calls, becomes less suited for these types of system and more data-centric query processing models are developed.

Furthermore, since analytic systems are designed to store large volumes of data, in-memory analytic DBMSs must carefully choose which parts of the dataset to keep in memory in order to minimize I/O reads at query time. The data structures used to store and index the in-memory data can also be substantially different from the ones traditionally used in clustered and secondary indexes in order to provide fast and cache-friendly scans as well as some form of data pruning. Lastly, transactions on these systems might also have different requirements due to the low number of transactions per second and high ingestion rates, creating opportunity for the development of novel and more lightweight concurrency control protocols.

This tutorial will discuss how in-memory analytic DBMSs are designed and built and outline the architecture of some state-of-art in-memory database systems, stressing the characteristics that differentiate them from the traditional DBMS design literature. In addition, the author will discuss some of the lessons learned while building and providing Cubrick as a service at Facebook, and highlight some of the many research opportunity avenues.

Author

Pedro Eugenio Rocha Pedreira is a Software Engineer at Facebook focused on database research. For the last 5 years he has led the team that develops Cubrick, a new in-memory analytic DBMS that targets interactive queries over highly dynamic datasets. Cubrick powers many internal analytic products and workloads and leverages a new indexing technique called Granular Partitioning, which he has proposed in his Ph.D thesis. Prior to joining Facebook, Pedro spent about 4 years working for the Brazilian Government, supporting the databases and infrastructure leveraged by the Brazilian Electoral Systems. He received his B.Sc., M.Sc and Ph.D from the Federal University of Parana (UFPR) in Curitiba/PR, Brazil.

References

Chen, J., Jindel, S., Walzer, R., Sen, R., Jimsheleishvilli, N., and Andrews, M. (2016). The MemSQL Query Optimizer: A modern optimizer for real-time analytics in a distributed database. *Proc. VLDB Endow.*, 9(13):1401–1412.

Diaconu, C., Freedman, C., Ismert, E., Larson, P.-A., Mittal, P., Stonecipher, R., Verma, N., and Zwilling, M. (2013). Hekaton: Sql server's memory-optimized oltp engine. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD'13, pages 1243–1254, New York, NY, USA. ACM.

Färber, F., Cha, S. K., Primsch, J., Bornhövd, C., Sigg, S., and Lehner, W. (2012). SAP HANA Database: Data Management for Modern Business Applications. *SIGMOD Rec.*, 40(4):45–51.

Kemper, A. and Neumann, T. (2011). HyPer: A Hybrid OLTP&OLAP Main Memory Database System Based on Virtual Memory Snapshots. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, ICDE '11, pages 195–206, Washington, DC, USA. IEEE Computer Society.

Kohn, A., Leis, V., and Neumann, T. (2018). Adaptive execution of compiled queries. In *Proceedings of the 2018 IEEE 34th International Conference on Data Engineering*, ICDE'18, Paris, France. IEEE Computer Society.

Mukherjee, N., Chavan, S., Colgan, M., Das, D., Gleeson, M., Hase, S., Holloway, A., Jin, H., Kamp, J., Kulkarni, K., Lahiri, T., Loaiza, J., Macnaughton, N., Marwah, V., Mullick, A., Witkowski, A., Yan, J., and Zait, M. (2015). Distributed architecture of oracle database in-memory. *Proc. VLDB Endow.*, 8(12):1630–1641.

Neumann, T. (2011). Efficiently compiling efficient query plans for modern hardware. *Proc. VLDB Endow.*, 4(9):539–550.

Pedreira, P., Croswhite, C., and Bona, L. (2016). Cubrick: Indexing millions of records per second for interactive analytics. *Proc. VLDB Endowment*, 9(13):1305–1316.

Pedreira, P., Lu, Y., Pershin, S., Dutta, A., and Croswhite, C. (2018). Rethinking concurrency control for in-memory olap dbmss. In *Proceedings of the 2018 IEEE 34th International Conference on Data Engineering*, ICDE'18, Paris, France. IEEE Computer Society.

Coleta, Integração e Pré-processamento de Dados de Múltiplas Fontes

Natércia A. Batista[1], Michele A. Brandão[1], Michele Brito[1], Daniel H. Dalip[2], Mirella M. Moro[1]

[1]Universidade Federal de Minas Gerais – Belo Horizonte – Brasil

[2]Centro Federal de Educação Tecnológica de Minas Gerais – Belo Horizonte – Brasil

{natercia,micheleabrandao,mibrito,mirella}@dcc.ufmg.br,
hasan@decom.cefetmg.br

Resumo. *Dados extraídos da Web são cada vez mais heterogêneos e não estruturados, representando desafios para atividades de coleta, integração e pré-processamento de dados. Existem estudos que são "orientados a dados", i.e., com base nos dados disponíveis, mas seus resultados ficam restritos aos respectivos dados. Em contraponto, vários problemas existem antes de se identificar quais dados são necessários para solucioná-los, e muitas vezes, são necessários dados de múltiplas fontes. Nesse contexto, o primeiro problema ao lidar com dados provenientes da Web é definir a estratégia de coleta, que pode ser classificada de acordo com o período e a forma de buscar a semente. Outro problema é definir uma estratégia para integrar os dados de diferentes fontes de forma a ter uma visão uniforme para usuários ou aplicações, além de armazená-los de maneira a permitir uma consulta eficiente. Finalmente, pode ser necessário realizar o pré-processamento de dados, que acontece antes ou depois da integração de dados, e envolve resolver dados faltantes e duplicados, normalização, etc. Este tutorial aborda esses três problemas de forma integrada com foco em questões práticas e de pesquisa.*

1. **Coleta de Dados.** Esta parte do tutorial cobre as principais estratégias de coleta de múltiplas fontes, bem como os principais desafios. Especificamente: uma visão geral sobre os três principais tipos de coleta considerando o período de realização da mesma (contínua, periódica e/ou ocasional) e a busca da semente -- entidade alvo de coleta ou ponto inicial (busca em largura, caminhamento aleatório e caminhamento aleatório Metropolis-Hastings [Gjoka et al. 2010]). Ademais, são apresentados exemplos práticos e desafios para cada estratégia.
2. **Estratégias para Integrar Dados de Múltiplas Fontes.** Após os dados serem coletados de múltiplas fontes, eles precisam ser integrados. Note que é possível coletar dados de cada fonte e armazená-los de forma separada para posterior integração, ou já armazenar todos os dados em um único local de forma integrada à medida que cada coleta é realizada. Nesta etapa, são abordadas vantagens e desvantagens dessas duas estratégias e das diferentes formas de armazenamento (planilhas, arquivos CSV, banco de dados relacionais, etc).

3. **Pré-processamento de Dados.** Nesta etapa do tutorial são abordados problemas geralmente encontrados nos dados após a realização da coleta: (*i*) valores faltantes - quando nenhum valor é armazenado para uma variável; (*ii*) veracidade dos dados - refere-se a viés, anormalidades e ruídos presentes nos dados; (*iii*) remoção de dados duplicados - dados iguais inseridos múltiplas vezes; (*iv*) falta de normalização - processo de reestruturar os dados a uma forma comum; e (*v*) redução dos dados - processo de minimizar a quantidade de dados a serem armazenados.
4. **Aplicações Reais.** Existem diferentes estudos que combinam múltiplas fontes de dados com objetivos distintos [Batista et al. 2017, Brandão et al. 2017, Dalip et al. 2013, Farnadi et al. 2018]. Por exemplo, Batista et al. (2017) medem a força dos relacionamentos entre desenvolvedores através da combinação de dados originados de um banco de dados relacional disponível na Web e dados coletados utilizando a API (*Application Programming Interface*) do GitHub. Brandão et al. (2017) utilizam dados de múltiplas fontes (DBLP e páginas Web) para fornecer visualizações com informações mais completas sobre pesquisadores. Ademais, Dalip et al. (2013) geram um ranking de respostas para perguntas no Stack Overflow por meio de indicadores. Dentre tais indicadores, utilizou-se a comparação entre os vocabulários das respostas com os vocabulários de bons artigos do Wikipedia. Todos esses exemplos de uso real da combinação de múltiplas fontes de dados são abordados nesta etapa mais prática do tutorial.
5. **Conclusão e Trabalhos Futuros.** Finalmente, além de abordar os principais conceitos, desafios e exemplos relacionados à coleta, integração e pré-processamento de dados de múltiplas fontes, discutimos problemas em aberto e possíveis trabalhos futuros a fim de incentivar pesquisas relacionadas a este tutorial.

Referências

Batista, N. A., Brandão, M. A., Alves, G. B., Silva, A. P. C. da, and Moro, M. M. (2017) "Collaboration strength metrics and analyses on GitHub." In *WI*, pp. 170-178.

Brandão, M. A., Diniz, M. A., Sousa, G. A., Moro, M. M. (2017) "Visualizing Co-Authorship Social Networks and Collaboration Recommendations With CNARe." *Graph Theoretic Approaches for Analyzing Large-Scale Social Networks*: 173-188.

Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2013) "Exploiting user feedback to learn to rank answers in q&a forums: a case study with stack overflow." In *SIGIR*, pp. 543-552.

Farnadi, G., Tang, J., Cock, M. D., and Moens, M-F. (2018) "User Profiling through Deep Multimodal Fusion." In *WSDM*, pp. 171-179.

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2010) "Walking in facebook: A case study of unbiased sampling of osns." In *IEEE Infocom*, pp. 1-9.

SBBD 2018

Keynotes

SBBD – Para Que e Para Quem?

Sérgio Lifschitz

Depto. de Informática – Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
Rua Marquês de São Vicente 225 – 22.451-900 – Rio de Janeiro – RJ – Brazil

sergio@inf.puc-rio.br

Resumo. *Neste ano de 2018, o 33o SBBD acontece junto com o 44o VLDB. O primeiro é o maior evento acadêmico nacional, enquanto o último é um dos maiores e mais importantes da área no âmbito internacional. Durante o SBBD, a comunidade brasileira de bancos de dados tem a oportunidade de acompanhar de perto os principais avanços científicos e tecnológicos na área, além de avaliar a importância do evento e o papel dos professores e pesquisadores na formação de pessoal e na contribuição para a sociedade. Nesta palestra, pretendo revisitar alguns dos diversos temas de P&D a que venho me dedicando ao longo dos anos, desde bancos de dados dedutivos até a análise de dados, passando pelas aplicações na bioinformática e pelos SGBDs autônomos. Aproveitarei também o convite para provocar uma reflexão sobre a nossa grande área de pesquisa, com foco na relevância e no papel de eventos como o SBBD para a ciência e a tecnologia no país.*

Sérgio Lifschitz is an Associate Professor at the Informatics Department of Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil. Sérgio has obtained his doctor's degree at the École Nationale Supérieure des Télécommunications (now Télécom ParisTech), Paris, France (1994). He also holds an M.Sc.(1990) and a B.Sc. (1986), both in Electrical Engineering, from PUC-Rio. His primary research area involves database systems, including applications in bioinformatics, social networks data, and autonomic tuning and management. Sérgio often participates in SBBD events, either with tutorials (three times), eleven full papers published, other 18 contributions for SBBD co-located events, one best paper and two best demo prizes, more than thirty times a member of evaluation committees and once the SBBD program chair.

Fast, Real-time Analysis on All Kinds of Data

Anastasia Ailamaki

Ecole Polytechnique Fédérale de Lausanne (EPFL)
BC 226, Station 14 – CH-1015 – Lausanne – Switzerland

anastasia.ailamaki@epfl.ch

Abstract. *Today's scientific and business processes heavily depend on fast and accurate data analysis. Data scientists are routinely overwhelmed by the effort needed to manage the volumes of data produced. As general-purpose data management software is often inefficient, hard to manage, or too generic to serve today's applications, businesses increasingly turn to specialised data management software, which can only handle one data format, and then resort to data integration solutions. With the exponential growth of dataset size and complexity, however, data format-specific solutions no longer scale for efficient analysis, thereby slowing down the cycle of analysing and understanding the data, and making decisions. I will illustrate the different nature of problems we face when managing heterogeneous datasets, and how these translate to fundamental challenges for the data management community. Then I will introduce RAW, a new solution inspired by these challenges. RAW overturns long-standing assumptions, enables meaningful and timely results, and promotes timely discovery.*

Anastasia Ailamaki is a Professor of Computer and Communication Sciences at the Ecole Polytechnique Federale de Lausanne (EPFL) in Switzerland and the co-founder of RAW Labs SA, a swiss company developing real-time analytics infrastructures for heterogeneous big data. Her research interests are in data-intensive systems and applications, and in particular (a) in strengthening the interaction between the database software and emerging hardware and I/O devices, and (b) in automating data management to support computationally- demanding, data-intensive scientific applications. She has received an ERC Consolidator Award (2013), a Finmeccanica endowed chair from the Computer Science Department at Carnegie Mellon (2007), a European Young Investigator Award from the European Science Foundation (2007), an Alfred P. Sloan Research Fellowship (2005), an NSF CAREER award (2002), and nine best-paper awards in database, storage, and computer architecture conferences,. She holds a Ph.D. in Computer Science from the University of Wisconsin-Madison in 2000. She is an ACM fellow, an IEEE fellow, the Laureate for the 2018 Nemitsas Prize in Computer Science, and an elected member of the Swiss National Research Council. She has served as a CRA-W mentor, and is a member of the Expert Network of the World Economic Forum.

Querying Graph Databases with the GSQL Query Language

Alin Deutsch

Dept Computer Science & Engineering – University of California, San Diego (UCSD)
9500 Gilman Drive – 92093-0404 – La Jolla – CA – USA

deutsch@cs.ucsd.edu

Abstract. *This talk presents GSQL, a recent addition to the spectrum of query languages for expressing graph analytics. GSQL is a high-level yet still Turing-complete language whose syntax is inspired by SQL in order to reduce the learning curve for SQL programmers, while simultaneously supporting a Map-Reduce interpretation that is preferred by NoSQL developers and that is conducive to massively parallel evaluation. The talk will also provide some context on the graph query language landscape represented in modern systems.*

Alin Deutsch is a professor of Computer Science and Engineering at UC San Diego. His research is motivated by the data management challenges raised by database-powered applications. Alin's interests include query language design and optimization for various data models ranging from text to the relational and post-relational models (with particular emphasis on graph data). He also works on cross-model data integration and on automatic verification of business processes. Alin earned his PhD in Computer Science from the University of Pennsylvania, an MSc degree from the Technical University of Darmstadt (Germany) and a BSc degree from the Polytechnic University Bucharest (Romania). He is the recipient of the 2018 ACM PODS Test of Time Award, a Jean D'Alembert Fellowship from the University Paris-Saclay, the Alfred P.Sloan Fellowship, the ACM SIGMOD 2006 Top-3 Best Paper Award, and an NSF CAREER award.

Reducing Errors by Refusing to Guess (Occasionally)

Dennis Shasha

Department of Computer Science– New York University (NYU)
251 Mercer Street – 10012 – New York – NY – USA

shasha@cs.nyu.edu

Abstract. *We propose a meta-algorithm to reduce the error rate of state-of-the-art machine learning algorithms by refusing to make predictions in certain cases even when the underlying algorithms suggest predictions. Intuitively, our SafePredict approach estimates the likelihood that a prediction will be in error and when that likelihood is high, the approach refuses to go along with that prediction. Unlike other approaches, we can probabilistically guarantee an error rate on predictions we do make (denoted the decisive predictions). Empirically on seven diverse data sets from genomics, ecology, image-recognition, and gaming, our method can probabilistically guarantee to reduce the error rate to 1/4 of what it is in the state-of-the-art machine learning algorithm at a cost of between 11% and 58% refusals. Competing state-of-the-art methods refuse at roughly twice the rate of ours (sometimes refusing all suggested predictions).*

Dennis Shasha is a Julius Silver Professor of computer science at the Courant Institute of New York University (NYU) and an Associate Director of NYU Wireless. He works on meta-algorithms for machine learning to achieve guaranteed correctness rates, with biologists on pattern discovery for network inference; with computational chemists on algorithms for protein design; with physicists and financial people on algorithms for time series; on clocked computation for DNA computing; and on computational reproducibility. Other areas of interest include database tuning as well as tree and graph matching. Because he likes to type, he has written six books of puzzles about a mathematical detective named Dr. Ecco, a biography about great computer scientists, and a book about the future of computing. He has also written five technical books about database tuning, biological pattern recognition, time series, DNA computing, resampling statistics, and causal inference in molecular networks. He has co-authored over eighty journal papers, seventy conference papers, and twenty-five patents. He has written the puzzle column for various publications including Scientific American, Dr. Dobb's Journal, and the Communications of the ACM. He is a fellow of the ACM and an INRIA International Chair.

Author Index