A Self-Portrayal of GI Junior Fellow Rainer Gemulla: Data Analysis at Scale

Rainer Gemulla

Abstract: My research focuses on methods to analyze and mine large datasets as well as their practical realizations and applications. The key question of interest to me is: How can we effectively and efficiently distill useful information from large, complex, and potentially noisy datasets? To approach this question, we are developing systems for scalable data analysis and data mining, for working with incomplete and noisy data, for data-intensive optimization, as well as for extracting structured information from natural-language text. This article highlights some of my work in these areas.

ACM CCS: Information systems → Information systems → Data mining; Computing methodologies → Parallel computing methodologies

Keywords: Data Analysis, Data Mining, Parallel Algorithms, Information Systems

1 Introduction

As more and more of the world's data becomes digitized and storage costs dwindle, information management systems are faced with new, unprecedented challenges. The ability to deeply analyze and understand the available data is becoming crucial in science and industry, and often requires the combination of techniques from areas such as database systems, distributed computing, data mining and machine learning, optimization, approximation techniques, or natural-language processing. This article gives a brief overview of my research on scalable data analysis and highlights some of the results we have obtained.

2 Approximation Techniques

The perhaps simplest way to handle large amounts of data is to not look at all of the data. Since my time as a PhD student [7], I am interested in techniques for approximate query processing in large, potentially distributed databases; such techniques produce high-quality approximate results in a fraction of the time required to compute the exact result. Approximation plays an integral role in data analysis; e.g., selectivity estimation techniques form the basis of query optimization in relational databases, and data samples are often used by analysts to facilitate interactive data exploration and mining.

My research focuses on theoretical and practical questions around building small representations of (aspects of) large databases, including samples, sketches, and synopses. We have developed methods that can create, exploit, combine, and transparently maintain such database samples and synopses. The resulting algorithms are efficient, effective, provably correct, and simple to implement. For example, the *random pairing* [9] (selected as one of the best papers of the renowned VLDB conference) and *augmented Bernoulli sampling* [10] algorithms maintain uniform random samples of a large database in the presence of insertion, update, and deletion transactions. Other examples include our work on *KMV synopses* [1] (selected as a research highlight for the Communications of the ACM) for distinct-value estimation as well our recent work that uncovers and corrects issues in existing sampling methods [8].

3 Data Analysis and Data Mining

A key area of my current research lies in methods and systems for data analysis and data mining. Aspects of my work include the support of complex analysis tasks (so-called *deep analytics*), simplicity of use, processing of semi-structured, incomplete, and noisy data, as well as scalable processing of very large datasets.

Jointly with researchers from the IBM Almaden Research Center, I have developed *Jaql* [2], a declarative scripting language for analyzing large semi-structured datasets in parallel using Hadoop's MapReduce framework. Jaql is a part of the IBM InfoSphere BigInsights and Cognos Consumer Insights products. The following

(a) An unknown data matrix (b) Known incomplete matrix (c) Automatically completed matrix

Figure 1: Matrix factorization with incomplete data

Jaql query, for example, outputs the annual average of a set of movie ratings:

```
read("ratings")
→ group by year = $.movie.year
    into { year, mean: avg($[*].rating) }
→ sort by [ $.year ]
```

With *Ricardo* [3], we have seamlessly integrated Jaql and the statistical software package *R*. Our methods allow usage of the sophisticated data analysis capabilities of R on large-scale, distributed datasets in a user-friendly way. To improve Hadoop's efficiency, we developed *CoHadoop* [6], a light-weight extension of Hadoop that improves efficiency by customized data placement strategies.

I am also interested in algorithms, in particular in methods that can handle incomplete and noisy datasets. One of the methods I worked on in this area is the factorization of large, potentially incomplete matrices. Matrix factorizations are the Swiss army knife of data mining, but many existing methods—such as the singular value decomposition—can handle incomplete data in a limited way or not at all. We have developed methods that can factorize matrices with millions of rows and columns and billions of known entries [11, 14]. Such factorizations are effective tools to "complete" incomplete matrices; see Fig. 1 for a visual example. Our work has been awarded with the IBM 2011 Pat Goldberg Memorial Best Paper Award in CS, EE, Math, the best paper award of the NIPS Big Learning workshops, and was selected as one of the best papers of the IEEE ICDM conference. In the context of recommender systems, in which matrix factorizations have been successfully employed to predict which users like which items, we have developed scalable matching techniques [12] that generate item recommendations under side constraints such as limited availability.

4 Text Mining

Another area of (more recent) interest to me is text mining and natural-language processing; most of my work in these area is conducted jointly with researchers from the Max Planck Institute for Computer Science, Germany. Our work is supported by a Google Focused Research Award and ultimately aims to analyze, structure, understand, and enrich large document collections in a way that allows effective and interactive search, exploration, and question answering.

We have developed *ClausIE* [4], a system that automatically extracts structured propositions from English sentences.[1] ClausIE is an open information extractor, i.e., extraction is performed in a domain-independent way. For instance, ClausIE extracts from the sentence

> "*Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.*"

the propositions

(*Bell*, *is*, *a telecommunication company*),
(*Bell*, *is based*, *in Los Angeles*),
(*Bell*, *makes*, *electronic products*),
(*Bell*, *makes*, *computer products*),
(*Bell*, *makes*, *building products*),
(*Bell*, *distributes*, *electronic products*),
(*Bell*, *distributes*, *computer products*),
(*Bell*, *distributes*, *building products*).

The resulting propositions are not disambiguated, i.e., we neither know that *Bell* refers to a company nor to which company. Our *Werdy* system [5] disambiguates word senses with a particular focus on verbs and can be used to enrich ClausIE's propositions.

Another line of my research in this area is pattern mining in sequential data, which includes textual data but also shopping transactions, error logs, and other kinds of event sequences. The simplest such patterns are perhaps n-grams, i.e., consecutive sequences of words or events that appear frequently. I aim to study methods that go significantly beyond n-grams in a number of important ways. Our *MG-FSM* system [13] is a first step in this direction: it performs pattern mining with positional or temporal "gaps." For example, our system discovered in the Netflix data that people frequently watch "Man in

[1] See `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/software/clausie/` for an online demo.

Black II", at most 1 day later "Independence Day," and at most another day later "I, Robot." MG-FSM scales to very large databases with billions of sequences.

5 Outlook

I have recently joined the Data and Web Science Research Group at the University of Mannheim, Germany, a successful group with a strong track record. If you are interested in performing research in this area, or if you need help from experts, we would be happy to talk to you.

Literature

[1] K. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla. On synopses for distinct-value estimation under multiset operations. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pages 199–210, 2007.

[2] K. S. Beyer, V. Ercegovac, R. Gemulla, A. Balmin, M. Eltabakh, C.-C. Kanne, F. Ozcan, and E. J. Shekita. Jaql: A scripting language for large scale semistructured data analysis. *Proc. VLDB Endowment*, 4(12):1272–1283, 2011.

[3] S. Das, Y. Sismanis, K. S. Beyer, R. Gemulla, P. J. Haas, and J. McPherson. Ricardo: Integrating R and Hadoop. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pages 987–998, 2010.

[4] L. Del Corro and R. Gemulla. ClausIE: Clause-based open information extraction. In *Proc. WWW Intl. Conf. on World Wide Web*, pages 355–366, 2013.

[5] L. Del Corro, R. Gemulla, and G. Weikum. Werdy: Recognition and disambiguation of verbs and verb phrases with syntactic and semantic pruning. In *To appear at Proc. EMNLP Conf. on Empirical Methods on Natural Language Processing*, 2014.

[6] M. Eltabakh, Y. Tian, F. Özcan, R. Gemulla, A. Krettek, and J. McPherson. CoHadoop: Flexible data placement and its exploitation in Hadoop. *Proc. VLDB Endowment*, 4:575–585, 2011.

[7] R. Gemulla. *Sampling algorithms for evolving datasets*. PhD thesis, Technische Universität Dresden, 2009. `http://nbn-resolving.de/urn:nbn:de:bsz:14-ds-1224861856184-11644`.

[8] R. Gemulla, P. J. Haas, and W. Lehner. Non-uniformity issues and workarounds in bounded-size sampling. *The VLDB Journal*, 22(6):753–772, 2013.

[9] R. Gemulla, W. Lehner, and P. J. Haas. A dip in the reservoir: Maintaining sample synopses of evolving datasets. In *Proc. VLDB Intl. Conf. on Very Large Data Bases*, pages 595–606, 2006.

[10] R. Gemulla, W. Lehner, and P. J. Haas. Maintaining Bernoulli samples over evolving multisets. In *Proc. ACM PODS Symp. on Principles of Database Systems*, pages 93–102, 2007.

[11] R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proc. ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 69–77, 2011.

[12] F. M. Manshadi, B. Awerbuch, R. Gemulla, R. Khandekar, J. Mestre, and M. Sozio. A distributed algorithm for large-scale generalized matching. *Proc. VLDB Endowment*, 6(9):613–624, 2013.

[13] I. Miliaraki, K. Berberich, R. Gemulla, and S. Zoupanos. Mind the gap: Large-scale frequent sequence mining. In *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pages 797–808, 2013.

[14] C. Teflioudi, F. Makari, and R. Gemulla. Distributed matrix completion. *Proc. IEEE ICDM Intl. Conf. on Data Mining*, 0:655–664, 2012.

Prof. Dr. Rainer Gemulla is a professor for computer science at the University of Mannheim, Germany, which he joined in 2014. He obtained his PhD from the Technical University of Dresden, Germany, in 2008, and worked subsequently as a postdoctoral researcher at the IBM Almaden Research Center in San Jose, CA, USA, and as a senior researcher at the Max Planck Institute for Informatics in Saarbrücken, Germany. Rainer's research focuses on data analysis and data mining techniques for efficiently extracting useful information from large, complex data collections. His work has been awarded with several awards, including multiple best-paper awards and a Google Focused Research Award. He serves as a junior fellow of the Gesellschaft für Informatik since September 2013.

Address: Universität Mannheim, Data and Web Science Research Group, 68131 Mannheim, Germany, Email: rgemulla@uni-mannheim.de