Research Interests – Alin Dobra

The recent explosion of the Internet and the rapid technological advances in gathering and storing information have resulted in huge amounts of data being collected at a very rapid rate. Developing ways to extract relevant information from such large amounts of data in a human comprehensible form and at the same time in a timely and cost effective way is of great practical importance.

Approximate Query Processing and *Data-mining*, both concerned with extracting useful knowledge from large amounts of data but using different premises, have been the subject of my research as a PhD candidate. My work, so far, is in most part of theoretical nature but the problems I attacked have direct practical applicability. For me theory is only a tool, albeit a very effective one for gaining interesting insights into the problem, but not an end in itself; I have always accompanied it by implementation and empirical validation. In what follows I will describe in some detail my particular interests in these two areas, pointing out past, current and future work.

Approximate Query Processing

The central problem in Data-mining is finding *interesting* patterns in the data in a scalable manner. Approximate query processing has a different premise namely that the user is interested in the answer of one or more *queries*, that can be expressed in SQL or in general as a mathematical formula, but an exact answer to such queries is not necessary as long as a reasonable guarantee of the precision of the result is provided. By allowing approximations, at least in principle, the computational effort and memory requirements of the algorithms can be drastically reduced.

Approximate Computation over Streaming Data, a subpart of approximate query processing, restricts the computational model by only allowing a single pass over the data in fixed (arbitrary) order and using only limited amounts of memory (usually logarithmic in the size of the stream). Computation under such restrictions is very interesting theoretically but at the same time very relevant for practical applications like network monitoring in large network installations, one pass query processing in databases and as a possible building block for Data-mining algorithms.

My current work in this area, in collaboration with Dr. Rajeev Rastogi and Dr. Minos Garofalakis from Bell-Labs and my thesis advisor Prof. Johannes Gehrke, focuses on the approximation of complex aggregate SQL queries that involve general equi-joins of multiple relations. Our main idea is to summarize the information in the streaming relations by random projections, which we call *sketches*, and to use these sketches to provide approximate answers to the aggregate queries together with approximation guarantees. Such sketches are easily maintainable over the streaming relations and, even more interestingly, due to their simplicity and linearity they are very suitable for performing highly distributed computations as is the case for querying sensor networks or large number of sources of data.

In addition to the basic sketching method we investigated ways to improve the quality of the approximation in the presence of extra information:

- **Sketch partitioning.** The main idea is to use existing coarse statistical information on the base data (e.g. histograms) to *intelligently* partition the domain of the underlying attributes. By applying the sketch method on the sub-problems obtained through such a partitioning and by subsequently combining the results we found that the error of the estimation is significantly reduced. We analyzed the optimization problems that arise and showed how good partitioning can be efficiently found in the general case.

- **Sketch sharing.** In the presence of multiple query expressions we showed how, by *intelligently* sharing sketches among the concurrent query evaluations, the approximation quality can be significantly improved. We also investigated the optimization problem of determining sketch sharing configurations that are optimal, under average and max error metrics, and provided both hardness results and practical heuristic/approximation algorithms.

Such techniques are important in practice since they provide a principled way to take advantage of available domain knowledge or prior information.

Data-mining

My main focus in Data-mining is on the efficient construction of *classification and regression trees.* These type of learners are expressive enough to compete well with other learners like support vector machines and neural networks. At the same time they are simple enough to be understood by users without much technical background and, more importantly from a research perspective, to be analyzed theoretically and to be constructed efficiently.

As part of my doctoral research at Cornell, under the supervision of Prof. Johannes Gehrke, I am studying three aspects of classification and regression tree construction:

- **Bias and bias correction in classification tree construction.** Often learning algorithms, in the presence of large amounts of noise, have undesirable preferences. In the case of classification and regression trees most methods for selecting the split variable have a strong preference for variables with large domains. In my work I provided a theoretical characterization of this preference and a general corrective method that can be applied to any split selection criteria to remove this undesirable bias. So far I have been able to apply this general method to a restricted setting, the correction of *gini gain* for discrete variables and k-ary splits.

- **Scalable linear regression tree construction**. In the presence of large amounts of data, efficiency of the learning algorithms with respect to the computational effort and memory requirements becomes very important. Part of my research is concerned with the scalable construction of regression trees with linear models in the leaves. The key to scalability is to use the EM Algorithm for Gaussian Mixtures to locally (at the level of each node being built) reduce the regression problem to a classification problem. As a side benefit, regression trees with oblique splits (involving a linear combination of predictor attributes instead of a single attribute) can be easily built.

- **Probabilistic classification and regression trees.** The use of strict split predicates in classification and regression trees has two undesirable consequences. First, data is fragmented at an exponential rate and therefore decisions in leaves are based on small number of samples. Second, decision boundaries are sharp because a single leaf is responsible for prediction. One principled way to address both these problems is to generalize classification and regression trees to make probabilistic decisions. More specifically, a probabilistic model is assigned to each branch and it is used to determine the probability to follow the branch. Instead of using a single leaf to predict the output for a given input, all leaves are used, but their contributions are weighted by the probability to reach them when starting from the root. The challenge is to find well motivated probabilistic models and to design scalable algorithms for building such probabilistic classification and regression trees.

Plans for Future Work

My current work provides interesting opportunities in both approximate query processing and data-mining. In the future I intend not only to enhance the applicability and usefulness of my initial results, but also to extend my research into related areas and to find novel applications for the theoretical developments. More precisely:

Approximate Query Processing

- **Extensions to other types of queries.** In the current form the sketch method applies only to computing COUNT and SUM aggregates over equi-joins without duplicate elimination. Extending the method to accommodate distinct values and more general queries is of great interest. Also combining the basic sketching technique with other types of synopses like sampling and histograms might allow for an extention of their applicability and for an increase of their performance.

- **Improvements of the basic scheme in the presence of extra knowledge.** So far we provided so far two ways to take advantage of such extra information: sketch partitioning and sketch sharing. Identifying other ways to use extra information in order to boost the accuracy of sketches while using the same amount of resources has great practical and theoretical importance. Promising types of prior information are, for example, (a) schema information: foreign keys and size of relations that will allow the combination of database technology with sketches, (b) statistical information – sketch partitioning uses such information but there are other ways, as some preliminary results suggest, statistical knowledge can be used to boost accuracy, and (c) workload information – our initial results on sketch sharing need to be extended when the sketch technique evolves.

- **Applications.** Even though we developed sketches for stream computation they have properties that make them a suitable summarization technique in other domains. Of particular interest are (a) highly distributed computation and querying sensor networks where sketches can allow approximations of complicated queries with very simple and decentralized algorithms, and (b) data-mining where approximate query processing can be used as a building block, for example in the fast computation of sufficient statistics necessary for classification tree construction. I plan to identify other application domains where sketches are useful and to build tools based on the theoretical developments that are useful in practice.

The theoretical results that we obtained in the process of developing the sketch techniques might be useful in other domains. In particular, the solutions to the optimization problems that arise might be relevant in other areas of computer science and possibly operations research. Also, our methods for incorporating domain knowledge can potentially be adapted to other application domains. In the long run I plan to investigate such connections in addition to developing novel approximation methods.

Data-mining

- **Extensions of the bias correction method.** In the near future I intend to develop a more complete theoretical characterization of the correction of the gini gain for k-ary splits, in order to show that none of the good properties of gini gain are removed. In the long run I want to attack the more technically difficult task of correcting gini gain for binary splits of both discrete and continuous variables, task that might require a combination of theoretical and empirical methods. In addition, such a development will provide insight into the statistical properties of gini gain, properties that I intend to use in other applications.

- **Scalable construction algorithms.** The use of the EM Algorithm for Gaussian mixtures was the key to obtaining a scalable algorithm for linear regression trees. I intend to further investigate how unsupervised learning in general can be used to speed up other data-mining techniques.

- **Probabilistic classification and regression trees** are in fact probabilistic models of the data. In future work I plan to investigate their usefulness to density estimation and inferring other types of information from the data.

Statistics proved to be an invaluable tool for attacking all these problems. I plan to continue identifying and exploiting connections between Data-mining, Machine Learning and Statistics in future work.

Hopefully, progress in these two areas will allow me to pursue my long term plan of seamlessly integrating approximate query processing, data-mining and relational database technology.