Efficient algorithms for collaborative decision making for large scale settings

Ira Assent
Department of Computer Science
Aarhus University
Denmark
ira@cs.au.dk

ABSTRACT

Collaborative decision making is a successful approach in settings where data analysis and querying can be done interactively. In large scale systems with huge data volumes or many users, collaboration is often hindered by impractical runtimes.

Existing work on improving collaboration focuses on avoiding redundancy for users working on the same task. While this improves the effectiveness of the user work process, the underlying query processing engine is typically considered a "black box"and left unchanged. Research in multiple query processing, on the other hand, ignores the application, and focuses on improving runtimes regardless of where the queries are issued from.

In this work, we claim that progress can be made by taking a novel, more holistic view of the problem. We discuss a new approach that combines the two strands of research on the user experience and query engine parts in order to bring about more effective and more efficient retrieval systems that support the users' decision making process.

We sketch promising research directions for more efficient algorithms for collaborative decision making, especially for large scale systems.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; H.2.4 [**Systems**]: Query Processing; H.2.8 [**Database Applications**]: Data mining

General Terms

Performance,Management,Algorithms

Keywords

Collaborative decision making, Interactive query processing

CIR'11, October 28, 2011, Glasgow, Scotland, UK.

1. INTRODUCTION

Collaboration has been recognized as an effective strategy in decision making. Joint analysis of data and comparison of results to related queries allows in-depth revision of available information and exchange of views that lead to well-founded decisions.

An important aspect in collaborative work on data is the interactive nature of the process. People working with the data expect near-real-time retrieval of data and processing of queries in order to experience a smooth working process that is not interrupted and delayed by response time of the underlying retrieval system.

On large data collections, and with many people submitting queries on related data, this poses a serious challenge to the query processing engine. As a consequence, collaborative decision making is limited to smaller data collections or to few participants.

In this paper, we propose addressing this efficiency issue by exploiting the fact that data and queries are typically related. Intelligent query processing techniques that are collaboration-aware can provide interactive data manipulation support where that is impossible using standard techniques.

This is different from techniques that study the relationship of queries at user interface level and different from multiple query processing. Both study just one side of the problem.

Here, we propose addressing them in sum in order to achieve the greatest benefit for interaction. We claim that this leads to improved interactivity and scalability of systems underlying collaborative decision making.

2. RELATED WORK

A groupware framework for collaborative work that supports development and testing of collaborative information retrieval strategies in search is proposed in [2]. The groupware architecture and source code facilitate the study of systems on the user's side, but do not target algorithm development that exploits collaborative work patterns.

The issue of dividing labor among collaborators during search is addressed in [4]. The goal is to reduce redundancy of search results among members of a team, and to support knowledge sharing among team members.

In [3], search results take the search results of fellow team members into account in order to fine-tune search for the group as a whole.

All of these techniques focus on improving the user experi-

ence in collaborative work by providing improved interaction and retrieval results.

This is in contrast to our proposal here: we claim that the user experience can be improved significantly by taking into account the algorithms underlying retrieval. By reducing computational overhead for search results that are potentially redundant or by re-using (partial) results that other team members have queried for, we can greatly improve the collaborative work experience. Results are detected more efficiently. This does not only lead to faster response times, but also enables interactive work with data and queries.

In query optimization, query processing has been made more efficient by optimizing the evaluation of several queries at the same time [7]. In these approaches, the aim is to derive query execution plans such that execution times are as low as possible.

For continuous queries, i.e., queries whose results are monitored and evaluated over changing data, related work exists for streaming data [1, 8]. Sensor networks are studied in [5].

Work on reducing the computational overhead by processing several related queries exists also for data mining tasks such as density-based clustering [9].

In our previous work, we study multiple queries in the context of similarity search on time series [6].

While this general concept of sharing computations among queries is widely used in different applications, these approaches are ignorant of the application that issues the queries. As a consequence, only queries that are issued at virtually the same time, are processed simultaneously, and past or future queries cannot be taken into account.

Also, the degree of similarity between queries is typically not known in advance. In collaborative decision making, by contrast, the general context of the queries is known, and some information about related queries could be made available to the query engine, as we propose in this work.

3. A SCENARIO: COLLABORATIVE DECISION MAKING & VISUALIZATION

The ideas discussed in this paper are developed as part of a project called "Improved decision making from massive data collections using wall-sized, highly interactive visualizations (WallViz)". The project aims to to explore the potential of interactive, wall-sized visualizations to support humans in their decision making process.

An overview illustration is given in Figure 1. Project members study interaction and visualization strategies, as well as data and information retrieval techniques for interactive settings. Our case partners come from different application areas, with a common interest in improving their decision making processes.

As part of the project, existing decision making processes are studied, and new collaborative approaches are developed. In this context, a variety of queries and data analysis tasks may be relevant. These range from simple data retrieval queries to "what-if" questions, from consideration of alternatives to complex data mining studies.

Common to these queries is that they are issued by human users who are directly involved in the decision making process. They are expected to submit a number of related retrieval queries, to compare, evaluate and commit to decisions.

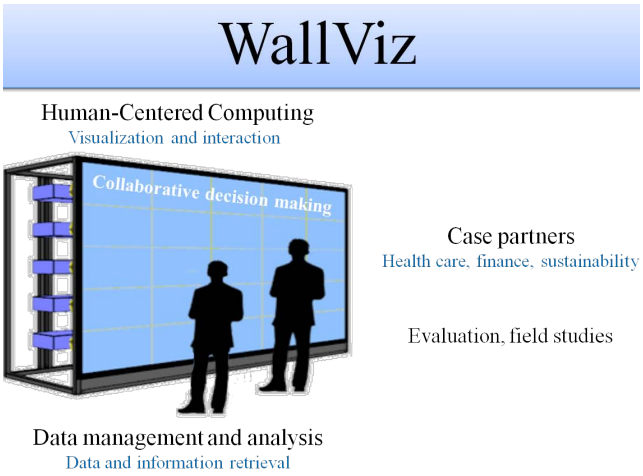

Figure 1: Overview over the WallViz project on collaborative decision making: Researchers in human-centered computing study visualization and interaction strategies, data management and analysis research focuses on data and information retrieval for interactive settings, and our case partners contribute real world applications.

4. COLLABORATIVE DECISION MAKING - USER EXPERIENCE & QUERY ENGINE

As discussed in the related work section, the user experience can be greatly improved by ensuring that retrieval results are collaboration-aware, i.e., that users do not receive redundant information that does not contribute to the overall knowledge of the group.

This redundancy removal does not necessarily translate to the best possible reduction in response time of the underlying retrieval algorithms. In the worst case, retrieval results are "cleaned" for redundant results after the algorithm has processed the same or very similar queries several times. Likewise, as mentioned above, multiple query processing strategies are successful in reducing the runtime once a set of queries is issued at nearly the same time.

Since these two strategies are employed in an isolated fashion, collaborative retrieval considers the query engine a "black box", and the query engine ignores the application that issues the queries. Consequently, the queries that are issued in order to improve the user experience are not necessarily optimized for the query engine. For example, documents could be filtered out or be re-ranked once the algorithm has completed its work.

Similarly, the query engine is not aware of other queries that might be issued in the near future, and might not even make full use of the ones that were processed in the past. In multiple query processing, the set of queries that exists at a given time is processed without considering future queries. This might lead to suboptimal query runtimes.

For example, if several different options are to be explored in a retrieval task, then processing a first batch of queries, and then a second batch might very well be slower than waiting for all queries before optimizing the query processing.

We claim that a holistic strategy can greatly improve the performance of systems that support collaborative decision making.

On the one hand, the query processing engine should be aware of the types of queries that are being issued in general, and of the time constraints that are associated with them in order to meet the requirements of an interactive work environment.

On the other hand, the optimization of the user experience should optimized the query distribution such that efficiency is improved. This leads to better runtimes and therewith to improved interactivity.

Our proposal for a more holistic approach can be summarized as: *know your past*, *know your present*, and *know your future*. What we mean by that is that there exist optimization possibilities depending on the time horizon of the queries that are being issued under collaborative work.

Know your past refers mainly to approaches that re-use or partially re-use query results. As collaborative work in decision making is likely to include very closely related queries and data, strategies such as caches, indexes on past results are important basics for better usage of resources.

Another traditional technique that could come into play here is the use of a "virtual past", i.e., materialization of queries that are likely to be used given knowledge about the decision task and / or typical work patterns of the users.

Know your present refers to collaboration-awareness at the time that redundancy-removal and other techniques for better user experience are applied: removing redundancy, for example, should be done in a manner that reduces the overhead for query processing at the same time.

Also, user specific information on preferences should be integrated directly, i.e., such that it can be used by the query processing engine in order to reduce the amount of data that needs to processed, instead of retrieving all possible results and cleaning or re-arranging them afterwards.

Finally, the most interesting aspect could be *know your future*. The more we know about typical user strategies in decision making, the better the query engine can prepare for upcoming queries. This not only refers to the above mentioned materialization and caching strategies, but most importantly for informed scheduling of several queries. More specifically, the query engine that is aware of future queries, might - provided time requirements allow this - wait for more queries before optimizing execution.

And, it could include queries that are not (yet) submitted to the system simply because they are very similar to the current ones. This leads to substantial efficiency gains when such queries are then posed at a later point.

The key notion in all of these approaches is the combination of the user experience view with the query engine's view: when optimizing each individually, we miss important speed-up possibilities. And such speed-up does not only lead to a more satisfactory experience for the users working with the system, but it makes interaction possible in scenarios where this is currently not possible simply because the response times of the system are far too high.

We believe that by bringing together researchers who are working on one of the two aspects, interesting new approaches for collaborative decision making can be developed.

5. CONCLUSION AND FUTURE WORK

In this work, we discuss improving interactive systems that support collaborative decision making.

A new paradigm is suggested that combines the successful work on better user experience in collaborative retrieval tasks with efficient query processing.

Rather than simply applying the techniques of the two, a holistic view on the problem is expected to lead to new algorithms that are collaboration-aware and highly interactive. As a consequence, collaboration incurs less redundancy for the user and the system.

In the recently started WallViz research project on collaborative decision making, we will study the workload on the query processing engine that is incurred by existing collaborative retrieval approaches, and we will extract profiles of the most efficient strategies for goals such as redundancy-removal.

We plan to develop instances of algorithms for specific data analysis and retrieval tasks that are collaboration-aware and that support users in their collaborative decision making process. We will analyze the improvement in interaction, and the potential for such algorithms in general. We will empirically study the scalability of these techniques for large scale systems in terms of data volume and number of users and of queries.

We are convinced that while algorithms for one type of data do not work exactly the same for other data types, strategies will evolve that make it possible to create algorithms for other data types as well.

As an example, we could imagine an algorithm that optimizes the near future by maintaining information on issued queries or (partial) results. It might build on a measure of similarity of queries or results and on a measure for the expected usefulness of storing this information. While such measures would most likely need to be defined with respect to the data type, such as simple measurements or complex multimedia content, as well as with respect to the decision making context, the algorithmic strategy itself is likely to generalize to other applications.

In this sense, a holistic paradigm is expected to lead to a set of algorithmic tools for collaborative retrieval tasks in general.

6. ACKNOWLEDGMENTS

This work has been supported in part by the Danish Council for Strategic Research, grant 10-092316.

We would like to thank Kasper Hornbæk, University of Copenhagen, for the original version of the WallViz figure.

7. REFERENCES

[1] S. Babu and J. Widom. Continuous queries over data streams. *ACM Special Interest Group on Management of Data (SIGMOD) Record*, 30(3):109–120, 2001.

[2] J. M. Fernández-Luna, J. F. Huete, R. Pérez-Vázquez, and J. C. Rodrīguez-Cano. Cirlab: A groupware framework for collaborative information retrieval research. *Information Processing & Management*, 46(6):749 – 761, 2009.

[3] C. Foley and A. Smeaton. Synchronous collaborative information retrieval: Techniques and evaluation. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 42–53. Springer Berlin / Heidelberg, 2009.

[4] C. Foley, A. Smeaton, and H. Lee. Synchronous collaborative information retrieval with relevance feedback. In *Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pages 1–4. IEEE, 2006.

[5] O. Jurca, S. Michel, A. Herrmann, and K. Aberer. Continuous query evaluation over distributed sensor networks. In *Proceedings of the IEEE 26th International Conference on Data Engineering (ICDE)*, pages 912–923. IEEE, 2010.

[6] H. Kremer, S. Günnemann, A. M. Ivanescu, I. Assent, and T. Seidl. Efficient processing of multiple dtw queries in time series databases. In *Proceedings of the 23nd International Conference on Scientific and Statistical Database Management (SSDBM 2011), Portland, Oregon, USA*, Heidelberg, Germany, 2011. Springer.

[7] T. Sellis. Multiple-query optimization. *ACM Transactions on Database Systems (TODS)*, 13(1):23–52, 1988.

[8] W. H. Tok and S. Bressan. Efficient and adaptive processing of multiple continuous queries. In *EDBT*, pages 215–232. Springer, 2002.

[9] D. Yang, E. A. Rundensteiner, and M. O. Ward. A shared execution strategy for multiple pattern mining requests over streaming data. *Proceedings of the VLDB Endowment*, 2(1):874–885, 2009.