Lecture Notes in Computer Science 3128

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Dmitri Asonov

Querying Databases Privately

A New Approach to Private Information Retrieval

Springer

Author

Dmitri Asonov
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95123, USA
E-mail: dasonov@us.ibm.com

Dissertation der Humboldt Universität zu Berlin
Tag der mündlichen Prüfung: 7. Juli 2003

Referent: Prof. Johann-Christoph Freytag, Ph.D., Humboldt Universität zu Berlin
Referent: Prof. Oliver Günther, Ph.D., Humboldt Universität zu Berlin
Referent: Rakesh Agrawal, Ph.D., IBM Almaden Research Center

Library of Congress Control Number: 2004094685

CR Subject Classification (1998): H.3, H.2, H.4, K.4, K.6.5, C.2

ISSN 0302-9743
ISBN 3-540-22441-6 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11018001 06/3142 5 4 3 2 1 0

Foreword

The Internet and the World Wide Web (WWW) play an increasingly important role in our today's activities. More and more we use the Web to buy goods and to inform ourselves about cultural, political, economical, medical, and scientific developments. For example, accessing flight schedules, medical data, or retrieving stock information become common practice in today's world. Many people assume that there is no one who "watches" them when accessing this data.

However, sensitive users who access electronic shops (e-shops) might have observed that this assumption often is not true. In many cases, E-shops track the users' "access behavior" when browsing the Web pages of the e-shop thus deriving "access patterns" for individual shoppers. Therefore, this knowledge on access behavior and access patters allows the system to tailor access to Web pages for that user to his/her specific needs in the future. This tracking of users might be considered harmless and "acceptable" in many cases. However, in cases when this information is used to harm a person - for example about the person's health problems - or to violate his/her privacy (for example finding out about his/her financial situation), he/she would like to be sure that such tracking is impossible to protect the user's rights.

These simple examples clearly demonstrate the necessity to shield the user from such spying **to protect his/her privacy**. That is, a user should be able to access a database (or a data source in general) without allowing others to "observe" which data is requested and accessed by the user; neither the query nor the answer should be visible or accessible to others. Surprisingly, despite the urgent need for concepts and techniques to protect the user from being spied on, very few results are known and available that addresses the problem adequately. During the last 10 years the area of **Private Information Retrieval (PIR)** has addressed some of the problems concerning privacy. However many of those results are of theoretical nature and thus do not carry over into practical solutions for protecting privacy when accessing information sources on the Web or in databases.

With this book Dr. Asonov is one of the first researchers who addresses the topic of querying data privately in a systematic and comprehensive way developing practical solutions in the context of database systems. The results presented in this book sometimes might look theoretical, but they describe

his clear understanding of the problem as well as the solutions required for "real-world" settings, in particular for scalable database solutions. As a basis Dr. Asonov first presents the framework for privately accessing database by developing several algorithms which also include the use of special hardware. In the second part of the book he focuses on solving several important subproblems; for them he also includes some validation by benchmarking to show to efficiency of the solutions. Finally, Dr. Asonov shows how his solutions could be used in solving some problems in the area of voting and digital rights management. Initially, these problems seem to be completely unrelated to PIR, however Dr. Asonov shows how some of his results can be used for creative solutions in the areas mentioned. Overall, the careful reader will notice that - despite the many technical details -his in-depth treatment of privacy in database provides the insight into the problem necessary for such an important topic.

In summary, with this book Dr. Asonov provides a systematic treatment of the problem how to access databases privately. The way he approaches the problem and how he develops solutions makes this book valuable for both researchers and practitioners who are interested in better understanding the issues. He develops scalable solutions that are necessary and important in the context of private information retrieval/private database access. The in-depth presentation of the algorithms and techniques is enlightening to students and a valuable resource for computer scientists. I predict that this book will provide the "starting point" for others to perform further research and development in this area.

Prof. Johann-Christoph Freytag, Ph.D., May 2004

Preface

People often retrieve information by querying databases. Designing databases that allow a user to execute queries *efficiently* is a subject that has been investigated for decades, and is now often regarded as a "researched to death" topic. However, the evolution of information technologies and society makes the database area a consistent source of new, previously unimaginable research challenges. This work is dedicated to partially meeting one of these new challenges: querying databases *privately*.

This new challenge is due to a very fundamental constraint of the conventional concept of querying information. Namely, in the conventional setting, the one who queries (the user) must reveal the query content and, by implication, the result of querying to the one who processes the query (the database server). This constraint seems to be negligible if the user trusts the server. However, the growing population of information providers makes it extremely difficult for users to establish and rely on the trustworthiness of information providers. Indeed, more and more cases are reported wherein information providers misuse the information provided by users' queries against the users, for example by sharing this information with third parties without permission, or by using this information for unsolicited advertisement.

We approach this constraint in a direct manner: If it is difficult to trust the server, we could try to remove the need for trust completely, by hiding the content of the user query and result from the server. This research problem, called Private Information Retrieval (PIR), has been under intensive and mainly theoretical investigation since 1996. These results are classified and analyzed in the first of four parts of this book. Our main contribution is considering this problem from a practical angle, as follows.

In Part II, we accept the assumptions and simplifications made in previous related work, and focus on obtaining efficient solutions and algorithms without changing the common model. Namely, we break the established belief that the server must read the entire database for a PIR protocol to answer a query. We further develop our solution by improving the processing and preprocessing complexities of our PIR protocol.

In Part III we extend the common PIR model in two directions. First, we relax the requirement that no information about a query must be revealed. This allows us to offer the user a trade-off between the level of privacy required

and the response time for a query. The second extension of the model is done by understanding the economics associated with the PIR problem. Namely, we assumed that information in the database is from different owners. We then consider the problem of distributing royalties between the information owners, given that no information about content of user queries is revealed.

A number of questions remain to be answered before the problem of querying databases privately can be regarded as completely investigated. However, we argue that results presented in the book have pushed the state of the art in this area, from the entirely theoretical level to the stage where implementing an applicable prototype can be considered ultimately possible.

Acknowledgements

I am most indebted to Professor Johann-Christoph Freytag for the success of this work. Our interaction was an example of a brilliant collaboration between a student and an adviser, so rarely found in science.

I was lucky to secure Professor Oliver Günther as my second advisor. I learned a lot from him. Professor Günther naturally supplemented the image of a perfect professor that I perceived from my first advisor.

I am very grateful to Rakesh Agrawal from IBM Almaden Research Center for being an external reviewer of my dissertation. Professor Sean W. Smith and Alex Iliev from Dartmouth College, Ronald Perez from IBM T. J. Watson Research Center, Christian Cachin from IBM Zürich Research Laboratory, and Frank Leymann from IBM Laboratory Böblingen were my occasional, but nevertheless most valuable external contacts.

I could not survive the hardship of making a Ph.D. without the warm, social support from my graduate school colleagues, and the team of DBIS department of Humboldt University. Especially, I would like to thank Markus Schaal and Christoph Hartwich for our fruitful collaboration in CS research, and my officemates Felix Naumann and Heiko Müller, who had to listen to my erroneous German every day. Ulrike Scholz and Heinz Werner have made DBIS a very comfortable place to work at.

My russian–speaking friends in Berlin, Stanislav Isaenko, Viktor Malyarchuk, and Mykhaylo Semtsiv helped me better understand research as a process by sharing their experiences in biological and physical research.

My teachers in Moscow provided the educational background from which I am benefiting now. Among them Yulia A. Azovzeva, Alexei I. Belousov, Valeri M. Chernenki, Maria T. Lepeshkina, Sergei V. Nesterov, Valentina P. Strekalova, Sergei A. Trofimov, and Valeri D. Vurdov were most helpful.

Last but not least, I am thankful to my family who supported me all the way through.

This research was supported by the German Research Society, Berlin-Brandenburg Graduate School in Distributed Information Systems (DFG grants no. GRK 316 and GRK 316/2).

Table of Contents