SIGMOD Record – September 2006 – Volume 35 Issue 3 –

SIGMOD OFFICERS, COMMITTEES AND AWARDS 1
EDITOR'S NOTES 2
CHAIR'S MESSAGE 3

[**Editor's note**: With the exception of the last pages –which would be the back cover of the printed issue– that are not included in this file, it has the same contents as the printed edition. All the articles are also available individually online and have been put together here for convenience only.]

SIGMOD Officers, Committees, and Awardees

Chair
Raghu Ramakrishnan
Department of Computer Sciences
University of Wisconsin-Madison
1210 West Dayton Street
Madison, WI 53706-1685
USA
raghu@cs.wisc.edu

Vice-Chair
Yannis Ioannidis
University Of Athens
Department of Informatics & Telecom
Panepistimioupolis, Informatics Bldngs
157 84 Ilissia, Athens
HELLAS
yannis@di.uoa.gr

Secretary/Treasurer
Mary Fernández
ATT Labs - Research
180 Park Ave., Bldg 103, E277
Florham Park, NJ 07932-0971
USA
mff@research.att.com

Information Director: Alexandros Labrinidis, University of Pittsburgh, labrinid@cs.pitt.edu.

Associate Information Directors: Marcelo Arenas, Ugur Cetintemel, Manfred Jeusfeld, Dongwon Lee, Michael Ley, Rachel Pottinger, Altigran Soares da Silva and Jun Yang.

Advisory Board: Tamer Ozsu (Chair), University of Waterloo, tozsu@cs.uwaterloo.ca, Rakesh Agrawal, Phil Bernstein, Peter Buneman, David DeWitt, Hector Garcia-Molina, Jim Gray, Masaru Kitsuregawa, Jiawei Han, Alberto Laender, Krithi Ramamritham, Hans Schek, Rick Snodgrass, and Gerhard Weikum.

SIGMOD Conference Coordinator: Jianwen Su, UC Santa Barbara, su@cs.ucsb.edu

SIGMOD Workshops Coordinator: Laurent Amsaleg, IRISA Lab, Laurent.Amsaleg@irisa.fr

SIGMOD Record Editorial Board: Mario A. Nascimento (Editor), University of Alberta, mn@cs.ualberta.ca, José Blakeley, Ugur Çetintemel, Brian Cooper, Andrew Eisenberg, Leonid Libkin, Alexandros Labrinidis, Jim Melton, Len Seligman, Jignesh Patel, Ken Ross, Marianne Winslett.

SIGMOD Anthology Editor: Curtis Dyreson (Editor), Washington State University, cdyreson@eecs.wsu.edu.

SIGMOD DiSC Editors: Shahram Ghandeharizadeh, USC, shahram@pollux.usc.edu and Joachim Hammer, UFL, jhammer@cise.ufl.edu.

PODS Executive: Phokion Kolaitis (Chair), IBM Almaden, kolaitis@almaden.ibm.com, Foto Afrati, Catriel Beeri, Georg Gottlob, Leonid Libkin, Jan Van Den Bussche

Sister Society Liaisons: Raghu Ramakhrishna (SIGKDD), Yannis Ioannidis (EDBT Endowment).

Awards Committee: Serge Abiteboul (Chair), INRIA, serge.abiteboul@inria.fr, Mike Carey, David Maier, Moshe Y. Vardi, Gerhard Weikum.

Award Recipients:

Innovation Award: Michael Stonebraker, Jim Gray, Philip Bernstein, David DeWitt, C. Mohan, David Maier, Serge Abiteboul, Hector Garcia-Molina, Rakesh Agrawal, Rudolf Bayer, Patricia Selinger, Don Chamberlin, Ronald Fagin, Michael Carey, and Jeffrey D. Ullman.

Contributions Award: Maria Zemankova, Gio Wiederhold, Yahiko Kambayashi, Jeffrey Ullman, Avi Silberschatz, Won Kim, Raghu Ramakrishnan, Laura Haas, Michael Carey, Daniel Rosenkrantz, Richard Snodgrass, Michael Ley, Surajit Chaudhuri, Hongjun Lu, and Tamer Özsu.

Editor Notes

As you probably noticed already, this issue is smaller than the recent ones, and I would like to make the reason for this the focus of these notes.

As you know there is a topic I have consistently addressed in recent issues of the *Record*. That is the one of the role of the *Record* in our community. After talking to many of our colleagues I am convinced that the *Record* is meant not to be "yet another journal" but instead fulfill its role as a high quality technical newsletter. As such it should contains articles that would not quite fit in a typical conference or workshop, granted, of course, those articles should still be mostly technical by nature. As an example of such articles I would refer, you to those analyzing database authorship and citations (e.g., two articles by Erhard and Thor, and by Sidiropoulos and Manolopoulos, both published in Dec./2005), and those discussing the single-blind vs. double-blind review (those by Madden and DeWitt, in the Jun./2006 issue, and two articles by Tung and Snodgrass, respectively, in this very issue). Needless to say the columns also play an adequate and important role in this scenario and have been handled very well thus far. (It is never enough to acknowledge and thank the volunteer help of the associate editors!)

Why am I saying all that you ask. In a sense to justify why this issue is, and likely the next ones to come will be, shorter. Once the view above is adopted many submitted papers, which would be otherwise worth publishing in the proceedings of a typical meeting, are no longer suitable for publication at the *Record*. As a consequence, the ratio of rejected papers has been increasing, thus leading to less "research articles" being published and finally resulting in shorter issues. This being said I still very much encourage submissions of technical papers with broader and/or provocative views, as well as comprehensive survey papers.

Another, orthogonal reason for this short issue, is that while I do have a good number of papers currently being under review, the reviewing process is taking longer than the usual. I ask the authors of papers which are waiting for the results of their submission to be patient. We all have to understand that peer-reviewing papers is a volunteer work. I have always tried to look for well-qualified reviewers, and typically those are the same people who are often recruited for the PC of good conferences. With conferences deadlines almost tied back-to-back, those people when not preparing a submission for a conference themselves are more often than not reviewing a conference submission by someone else. (Add to this the Summer, when most of us take some (deserved) time off.) Since conference reviews have tight deadlines it is not surprise (though not fortunate) that other reviews receive lower priority, hence taking longer to complete. Unfortunately I do not see an easy to solve this, though I am trying to get some commitment from reviewers I am also realistic about our workload and ever shifting priorities.

That is about what I wanted to say today. I hope you enjoy this issue, in particular the articles about our reviewing processes and their implications. I dare to suggest that our community (and not necessarily only ours) might want to do some (re)thinking about the role of conferences and reviewers. And I have to say that I am glad to see the *Record* being used to document such reflections. Cheers!

Mario Nascimento, Editor.
August, 2006.

Chair's Message

The main event to report on this time is the annual SIGMOD conference, held recently in Chicago. Seeing the organization up close, I'm amazed at the effort that goes into it, primarily from people who volunteer their time freely for the good of the community. We owe them a big round of applause, and sincere thanks:

The program committee chairs Surajit Chaudhuri (SIGMOD) and Jan van den Bussche (PODS) did a heroic job, together with the respective PCs.

Goce Trajcevski deserves special mention for his dedicated work on the thankless task of local arrangements. Peter Scheuermann did a great job of obtaining sponsorships, helping to secure the financial side of the conference, and I'd like to thank Clement Yu for taking on the demanding role of General Chair for the conference.

I'd like to take this opportunity to acknowledge the generous support provided by the following companies:

Oracle and Sybase (principal supporters); Google, IBM, Intel, Microsoft Research, and SAP (supporters); Hewlett Packard, Motorola, U. Hasselt, Yahoo! Research (contributors); Ask Jeeves (supporting organization)

A number of people gave generously of their time in a variety of capacities. Please see the following URL to see who they are, and how much work went into the conference!

http://tangra.si.umich.edu/clair/sigmod-pods06/

Sibel Adali and Vassilis Vassalos took on the task of selecting the SIGMOD Undergraduate Scholarship winners with very little notice, and did a great job, with the help of a committee assembled almost instantly: Torsten Grust, Sudipto Guha, Ihab Ilyas, Chen Li, Nikos Mamoulis, and Maria Esther Vidal. I just looked at the link above and noticed that they are not listed (an oversight because of the circumstances); none of them thought to point this out earlier, or perhaps, even to look. I think this underscores one of the strengths of our community … the willingness of busy researchers to take on voluntary tasks at short notice, for the good of the field rather than the recognition. I hope we never lose this spirit.

I'd like to give special thanks to the student volunteers:

Joel Booth, Hui Ding, Eduard Dragut, Fang Fang, Oliviu Ghica, Ali Hakim, Dongmei Jia, Ying Lai, Shuang Liu, Fang Liu, Ramanathan Narayanan, Berkin Ozisikyilmaz, Amira Rahal, Damian Roqueiro, Huiyong Xiao, Lin Xiao, Huabei Yin, Wei Zhang, and Wei Zhou.

Finally, as we approached the conference date, a number of issues surfaced and required considerable, and unanticipated, attention. I'd like to particularly thank Mary Fernandez, Ginger Ignatoff, and Joanne Martori for their help in this regard; of course, a number of other people involved in organizing the conference stepped up as well.

Turning to the main event, the conference highlighted a strong technical program with award-winning papers:

The SIGMOD Best Paper Award winner:

- Panagiotis Ipeirotis, Eugene Agichtein, Pranay Jain, Luis Gravano
  - To Search or to Crawl? Towards a Query Optimizer for Text-Centric Tasks

The SIGMOD Best Paper Award honourable mentions:

- Michalis Petropoulos, Alin Deutsch, Yannis Papakonstantinou
  - Interactive Query Formulation Over Web-Service Accessed Data Sources
- Izchak Sharfman, Assaf Schuster, Daniel Keren
  - A Geometric Approach to Monitoring Threshold Functions over Distributed Data Streams

The PODS Best Paper Award winner:

- Mikolaj Bojanczyk, Claire David, Anca Muscholl, Thomas Schwentick, Luc Segoufin
  - Two-Variable Logic on Data Trees and XML Reasoning

The PODS Best Newcomer Award winner:

- Michael Bender, Haodong Hu
  - An Adaptive Packed-Memory Array

In addition, several awards were made at this year's conference recognizing significant contributions over the years:

Jeff Ullman received the SIGMOD Edgar F. Codd Innovations Award.

Tamer Ozsu received the SIGMOD Contributions Award.

The SIGMOD Test-of-Time Award was presented jointly to:

- Tian Zhang, Raghu Ramakrishnan and Miron Livny
  - BIRCH: An Efficient Data Clustering Method for Very Large Databases
- Venky Harinarayan, Anand Rajaraman and Jeffrey D. Ullman
  - Implementing Data Cubes Efficiently

The SIGMOD Dissertation Awards Committee evaluated 18 submissions for the 2006 dissertation award. The main criteria for the evaluation were: Theory/Foundational Work, System Orientation and Impact. Congratulations to all the winners and their advisors; the pipeline of superb young researchers with their high standards and great enthusiasm is vital to our community's continued vibrance and impact:

The winner of the inaugural ACM SIGMOD Doctoral Dissertation Award is Gerome Miklau, who did his PhD at the Univ. of Washington, advised by Dan Suciu.

The two runners up are Marcelo Arenas (PhD from Univ. of Toronto; advisor Leonid Libkin) and Yanlei Diao (PhD from Univ. of California-Berkeley; advisor Michael Franklin).

Sincerely,
Raghu Ramakrishnan

Impact of Double Blind Reviewing on SIGMOD Publication: A More Detail Analysis

Anthony K. H. Tung
National University of Singapore
atung@comp.nus.edu.sg

1. Introduction

In [1], a set of statistic had been provided with the conclusion that double blind reviewing make no impact on SIGMOD publication. Our studies here will show results contrary to that finding.

2. Use of Median instead of Mean

Our first study will use the median instead of the mean for our analysis. The use of median is more robust to the existence of outliers which can skew the mean drastically [2]. For example, the mean of papers/famous person from 2001-2005 for SIGMOD is 0.912 while the median is 0.81. Also, 4 out of the 5 values used in the computation of the mean are in fact smaller than 0.912. Table 1 depict the data obtain from [1] with four additional rows at the end. Of these four additional rows, the first three are the column median for the period 1994-2005, 1994-2000 and 2001-2005 respectively. The last row computes the gain of the column median from the period 2001-2005 over the period 1994-2000. SIGMOD adopt double blind reviewing from the year 2001 onwards.

Table 1: Publication Statistics by Year Per Conference from [1].

	Papers/Famous Person		Fraction Famous Papers		Total Papers	
Year	SIGMOD	VLDB	SIGMOD	VLDB	SIGMOD	VLDB
1994	0.81	0.73	**0.48**	0.28	42	65
1995	0.54	0.73	0.37	0.31	36	59
1996	**0.88**	**1.07**	**0.47**	**0.55**	47	49
1997	**0.92**	0.85	**0.55**	**0.38**	42	55
1998	0.73	0.69	0.43	0.33	42	52
1999	**0.88**	0.81	**0.53**	**0.35**	42	58
2000	**1**	**0.88**	**0.52**	**0.38**	48	58
2001	0.77	0.81	0.44	0.31	44	66
2002	0.81	**1.15**	0.4	0.32	50	91
2003	**0.85**	**1.15**	0.4	**0.34**	53	84
2004	***1.34**	**1.53**	**0.49**	**0.38**	69	102
2005	0.81	**0.92**	0.31	0.22	66	103
Median(1994-2005)	0.83	0.865	0.455	0.335	45.5	62
Median(1994-2000) **A**	0.88	0.81	0.48	0.35	42	58
Median(2001-2005) **B**	0.81	1.15	0.4	0.32	53	91
(B-A)/A	-7.95%	+41.98%	-16.67%	-8.57%	+26.19%	+56.90%

Legends: **Red Bold**: Higher than Median(1994-2005) i.e. "good year"
* : Outlier

Based on Table 1, we have the following observations

Observation 1 (OB1): In term of total number of papers accepted, both SIGMOD and VLDB see an increase of 26.19% and 56.90% respectively for the period 2001-2005 when compared to the period 1994-2000.

Observation 2 (OB2): In term of papers/famous person, VLDB approximately follow the trend in (OB1) with a growth of 41.98%. However, this is not the case for SIGMOD which see a drop of 7.95%, a rather unexpected and surprising result considering (OB1) where the number of SIGMOD paper accepted increase by 26.19%.

Observation 3 (OB3): In term of fraction famous paper, both SIGMOD and VLDB see a drop indicating that the increase in accepted papers from (OB1) are not coming from the famous people. However, the drop for SIGMOD is 16.67% which is two times that of VLDB with a drop of only 8.57%.

Based on observations 1-3, we believe that there are indications that double blind reviewing does have an impact in term of papers accepted for famous people in SIGMOD.

3. Probability of "Good Year"

We first define a "good year" for famous person to be a year in which the value is higher than Median(1994-2005). For example, in term of fraction/famous person, the good years for famous person in SIGMOD are 1996, 1997, 1999, 2000, 2003 and 2004 since their corresponding values are all higher than 0.83. This is similar in spirit to the definition of "famous person" with a more compelling reason of ensuring that there are enough samples for both classes i.e. "good year" and "bad year".

Based on this definition, we will expect famous people to have 50% chance of having "good year" and "bad year" irregardless of DB or non-DB if they are in fact independent of each other. We have the following observations:

Observation 4 (OB4): In term of papers/famous person, we see that there is a 57% (i.e. 4/7) chance of having "good year" for SIGMOD from 1994-2000 while there is only a probability of 40%(i.e. 2/5) for having "good year" for SIGMOD from 2001-2005. This again is surprising given (OB1).

For VLDB, this trend is reversed. The chance of having "good year" for VLDB from 1994-2000 is 28.57%(2/7) while the chance of doing so from 2001-2005 is 80%(4/5). Comparing SIGMOD(2001-2005) to VLDB(2001-2005), the chance of having "good year" for SIGMOD is only half (40% vs 80%) that of VLDB which indicate that double blind reviewing halved the chance of "good year" for famous people.

Observation 5 (OB5): In term of fraction famous paper, both SIGMOD and VLDB see a drop in the probability of "good year" for 2001-2005. The explanation is similar to that of (OB3). However, the chance of "good year" for SIGMOD(2001-2005) is still only half that of VLDB(2001-2005) i.e. 20% vs 40%.

From observations 4-5, we can see that double blind review in fact reduce the probability of a "good year" for famous people by half, indicating its strong impact.

4. Final Conclusion

In this report, we make two studies which indicate that double blind review in SIGMOD do have its impact on the performance of "famous person" compared to VLDB. As mentioned in [1], there are probably a lot of other factors that must be taken into consideration before the database community makes a final choice on whether to continue with double blind review. We hope that our studies here can provide a little more information towards making the final choice.

Acknowledgement:

It is unfortunate that we have to disagree politely with the finding of Samuel and David in [1]. We nonetheless like to thank them for their effort in collecting the data, without which the study here will not be possible.

Reference:

1. Impact of Double-Blind Reviewing on SIGMOD Publication Rates. Samuel Madden and David Dewitt. SIGMOD Record, Volume 32, No. 2, June 2006.
2. http://www.ltcconline.net/greenl/courses/201/descstat/mean.htm

Single- Versus Double-Blind Reviewing: An Analysis of the Literature

Richard Snodgrass
rts@cs.arizona.edu

1 Introduction

The peer review process is generally acknowledged as central to the advancement of scholarly knowledge. It is also vital to the advancement of individual careers.

With so much at stake, it is important to examine, and re-examine, issues pertaining to review quality on an ongoing basis. Thus it is appropriate that controversy has arisen in our field pertaining to the practice of double-blind reviewing. "As scientists, we should rather welcome all occasions to reflect on the act of writing, evaluating, editing and publishing research findings. The issue of double-blind refereeing, which recurs periodically in scientific circles, provides us with such an opportunity" [Genest 1993, page 324].

Most database journals employ single-blind reviewing, in which the reviewer is unknown to the author, but the identity of the author is known to the reviewer. Others employ double-blind reviewing, in which the identity of the author and the reviewer are not known to each other. The arguments for double-blind reviewing are that it is fairer and that it produces higher quality reviews. The arguments advanced against double-blind reviewing include that it has little effect, that it makes it more difficult for reviewers to comprehensively judge the paper, and that it is onerous to administrate [Ceci & Peters 1984].

To shed light on this controversy, we examine the now substantial scholarly literature regarding blind reviewing. This literature includes empirical studies from biomedicine, communication, computer science, economics, education, medicine, public health, physics, and psychology, retrospective analyses from computer science, ecology, economics, and medicine, and a quantitative meta-analysis from psychology. It is useful and instructive to learn what other disciplines, using diverse approaches, have discovered about blind reviewing.

In the following, we first define the various terms used in the literature. We then examine in some depth the general issues of fairness, review quality, and efficacy of blinding. As will be seen, in most cases the results are mixed. We end with a list of recommendations from scholarly societies and a brief summary of this complex sociological question.

2 Terminology

ACM defines a *refereed journal* or *refereed conference* as one that "is subjected to a detailed peer review, following a defined, formal process according to a uniform set of criteria and standards." [1]. This is distinguished from *formally reviewed material* ("subjected to a structured evaluation and critique procedure following a defined process uniformly applied as with refereeing, only without requiring that the tests of scholarly originality, novelty and importance be applied"), *reviewed* ("subjected to a more informal and not necessarily uniform process of volunteer review, with standards dependent upon the publication and the type of material"), *highly edited* ("professionally edited, usually by paid staff, with primary emphasis on exposition, graphic presentation, and editorial style rather than on content and substance"), and *unreviewed* ("published as submitted, with or without copyediting"). "Reviewing" in the present document refers to peer review for a refereed journal or conference.

Peer review is the use of predetermined reviewers, in the case of program committees, or ad hoc reviewers, in the case of reviewers for most journals, who individually read the submitted manuscript and prepare a written review. Sometimes, as in the case of some conference program committees, reviewers will subsequently either physically or electronically meet to discuss the papers to arrive at an editorial decision. For most journals, the Associate Editor handling the paper or the Editor-in-Chief will make the final editorial decision.

In the vast majority of refereed database conferences and journals, the identity of the reviewer(s) is not revealed to the author(s), ostensibly to ensure more objective reviewing. This is termed *single-blind reviewing* or, less frequently, "one-eyed review" [Rosenblatt & Kirk 1980]. (Incidentally, the terms "reviewer" and "referee" are used interchangeably in the literature.)

There are other sources of confidentiality in the review process. For most journals, the identity of the reviewer is not revealed to other reviewers; such is not the case for program committees. Some conferences, such as IEEE ICDE, utilize area program chairs. Generally but not always it is known which area program chair mediated the editorial decision for a submission. The associate editor for a journal submission is usually revealed to the author, except when one of the authors is himself/herself an

[1] ACM Policy on Pre-Publication Evaluation, at http://www.acm.org/pubs/prepub_eval.html

associate editor (cf. the *TODS* policy [Snodgrass 2003]). Of course, the Editor-in-Chief and the Program Chair are known to everyone. The important point here is that the term "single-blind reviewing" applies only to hiding the identity of the reviewer from the author.

In an effort to achieve more objective reviewing, a venue can also request that the identity of the author be removed from the submitted manuscript, a process termed *blinding the manuscript*. When the identity of the authors and their institutions is kept from the reviewers, this is termed double-blind reviewing. Note that the Editor-in-Chief and Program Chair, and generally the Associate Editor, are made aware of this information via a separate cover sheet not shared with the reviewers.

The psychological sciences utilize a different terminology that conveys a subtle philosophical shift. When the identities of the authors and reviewers are not revealed to each other, it is termed in these sciences a *masked* reviewing process. Note the symmetry of this terminology. The American Psychological Association *Guide to Preparing Manuscripts for Journal Publication* [Calfee & Valencia 2006] states, "Peer review is the backbone of the review process. Most APA journals, like the majority of other professional publications, practice anonymous, or masked, reviews. Authors and reviewers are unaware of each other's identities in most instances, an arrangement designed to make the process more impartial." The implication is that revealing either the reviewer's identity or the author's identity breaks the mask. Presumably single-blind reviewing would then be termed "non-masked," but the APA doesn't use the term. (The term "unmasking" denotes revealing the identity of a reviewer to a co-reviewer [van Rooyen 1999]; we don't consider that practice here.)

The present paper will use the terms single-blind and double-blind reviewing, as well as their respective three-letter acronyms, *SBR* and *DBR*.

Venues differ in who does the blinding/masking of a submission. We will use the term "author masking" when the author removes identification from the paper before submitting it and "editorial masking" when such identification (generally, author name and affiliation) is removed in the editorial process before sending the manuscript to the reviewers. Procedures differ in how aggressive is the required author masking and the actual editorial masking. Self-citations and other first-person references in the body of papers are generally retained in editorial masking. While author masking can be more thorough, because authors would know what kind of information is revealing, authors through various devious means can circumvent both kinds of masking.

3 Literature Reviews

Because of the centrality of peer review to the propagation of scientific knowledge, one would expect that peer review has been thoroughly studied, with its benefits and potential pitfalls exhaustively documented. Such is not the case. Prior to 1975 research on peer review was relatively scarce, with discussions based more on personal observations rather than systematic data gathering. Campanario has written the most comprehensive (61 page!) summary of the research that has been done on peer review, generally over the two decades of 1975–1995. Part 1 of this summary covered the participants in the system: the credentials of referees, editorial board members, and editors; how editors and editorial board members are appointed; how referees are chosen; reviewer incentives and tasks; and systemic problems of reliability, accuracy, and bias: reliability of review; accuracy of review; is the system biased towards positive results; and is the system biased against replication [Campanario 1998a]. Part 2 covered current research findings about fraud, favoritism, and self-interest in peer review [Campanario 1998b]; this part included a three-page section on DBR.

The Institute of Mathematical Statistics (IMS) formed the Ad Hoc Committee on Double-Blind Refereeing in February 1991, at least partly in response to a report of the New Researchers Committee [Altman et al. 1991]. This committee issued a report that contained a three-page literature review [Cox et al. 1993].

Other reviews include one of the voluminous literature (over 600 items) on the more general topic of journal reviewing, including several paragraphs related to DBR [Dalton 1995], another on 68 papers on empirical evidence concerning journal peer review [Armstrong 1997], including one page on DBR, and a summary of the evidence for the effectiveness of peer review in general, including about a page and a half on DBR [Fletcher & Fletcher 1997]. Finally, two international congresses on editorial peer review have been held, with papers revised and re-reviewed and appearing in the *Journal of the American Medical Association (JAMA)* [Rennie 1990, Rennie & Flanagin 1994]. Relevant papers from these congresses are discussed in the following sections.

These literature reviews emphasize three primary aspects relevant to blind reviewing, fairness to authors (to unknown authors or to authors affiliated with unknown institutions, to less-published or to proficient authors, to both genders), review quality, and blinding efficacy. The following sections will address each aspect in turn.

4 Fairness

The fundamental argument for double-blind reviewing is that it is fairer to authors (and thus, indirectly, to readers). The argument proceeds as follows: The judgment of whether a paper should be accepted for publication should be made on the basis of the paper alone: is what the submission states correct, insightful, and an advancement of the state-of-the-art? The editorial judgment should not be made on extenuating circumstances such as who wrote the paper or the professional affiliations of the authors. By blinding the submission, the reviewers cannot take these peripheral aspects, which are not relevant, into account in their review.

The analysis of fairness in the extant literature concerns (a) fairness to unknown authors or institutions, (b) fairness to prolific or to less-published authors, and (c) gender equity. There is also the related issue of the perception of fairness. The following sections will elaborate on each of these concerns.

4.1 Fairness to Unknown Authors or Institutions

Some evidence from retrospective and experimental studies suggest that when the authors' names and affiliations are known, reviewers may be biased against papers from unknown authors or institutions, termed "status bias" [Cox et al. 1993]. An anecdote illustrates this possibility. The psychologist Robert Rosenthal wrote of his experience in the prestigious journal *Behavioral and Brain Sciences* with "the 15 to 20 articles I had written while at UND [University of North Dakota] that I was not able to publish in mainstream psychological journals. After I had been at Harvard a few years, most of those same articles *were* published in mainstream journals. My anecdote does not demonstrate that journal articles were biased against papers from UND and biased toward papers from Harvard. There are plausible rival hypotheses that cannot be ruled out. My belief, however, is that location status bias may well have played some role in the change in publishability of my stack of papers" [Rosenthal 1982, page 235].

We now examine the studies that attempt to detect status bias, in chronological order.

A retrospective study of manuscripts that had been submitted to *The Physical Review* between 1948 and 1956 found that "some 91 per cent. of the papers by physicists in the foremost departments were accepted as against 72 per cent. from other universities" [Zuckerman & Merton 1971, page 85]. Two possible explanations were offered: status bias and "differences in the scientific quality of the manuscripts coming from different sources" [ibid].

An early experiment found that "the effect of institutional prestige failed to attain significance in any one of the measures" [Mahoney et al. 1978, page 70]. "Experimental manuscripts were sent to 68 volunteer reviewers from two behavioristic journals. ... Institutional affiliation was also manipulated on the experimental manuscripts, with half allegedly emanating from a prestigious university or a relatively unknown college" [ibid].

Another retrospective study, this of the records of reviews of a society which publishes research journals in two areas of the physical sciences, found large differences in how papers from minor and major universities are reviewed: "minor university authors are more frequently evaluated favourably (ie less critically) by minor university referees, while major university authors are more often evaluated favourably by major university referees than they are by those affiliated to minor universities. It would therefore appear that when referees and authors in these areas of the physical sciences share membership of national or institutional groups, the chances that the referees will be less critical are increased. ... Personal ties and extra-scientific preferences and prejudices might, of course, be playing a part as well. But it appears that, even in the absence of these personal factors, the scientific predispositions of referees still bias them towards less critical evaluation of colleagues who come from similar institutional or national groups, and so share to a greater extent sets of beliefs on what constitutes good research" [Gordon 1980, pp. 274–5].

Peters and Ceci performed a famous experiment [Peters & Ceci 1982] that gave some credence to the existence of such bias.[2] In this study, twelve papers published by investigators from prestigious and highly productive American psychology departments in high-quality journals were altered with fictitious names and institutions substituted for the original ones and then formally resubmitted to the journals that had originally refereed and published them 18 to 32 months earlier. Only three were detected as resubmissions; of the remaining nine, eight were rejected, in many cases based on "serious methodological flaws."

Peters and Ceci put forth the possibility of status bias: "The predominantly negative evaluations of the resubmissions may reflect some form of response bias in favor of the original authors as a function of their association with prestigious institutions. These individuals may have received a less critical, more benign evaluation than did our unknown authors from "no-name" institutions. ... The near perfect reviewer agreement regarding

[2] This experiment and the associated paper have generated much controversy. A special issue of *Behavioral and Brain Science* was dedicated to the paper and 55 (!) commentaries, along with an authors' response that was almost as long as the original paper. "In the course of the Commentary, just about every aspect of the peer-review problem is brought up and subjected to critical scrutiny" [Harnad 1982, page 186].

the unacceptability of the resubmitted manuscripts, coupled with the presumably near perfect agreement among the original reviewers in favor of publishing, provide additional convergent support for the response bias hypothesis" [Peters & Ceci 1982, page 192]. Their proposed solution: "If institutional affiliation or professional status can in fact bias peer review - and this bias proves to have no validity, or negative validity - then one possible solution to this problem (as several critics have recommended) would be to establish blind reviews as standard journal policy" [ibid, page 194].

A seminal experiment [Blank 1991] demonstrated status bias in reviewing more directly. In this experiment, every other paper that arrived at the *American Economic Review* was designated as double-blind. For these papers, an editorial assistant removed the name and affiliation of the author from the title page and typically scanned the first page for additional titles or notes that would identify the author (i.e., editorial masking). This experiment lasted for two years.

The relevant issue was "whether the ratio of acceptance rates between institutional ranks in the blind sample differs from the corresponding ratio in the nonblind sample." [Blank 1991, page 1053–1054]. It was found that this ratio did not differ for those at top-ranked departments and those at colleges and low-ranked universities. All other groups, in that important gray area where editorial judgment is most needed, had substantially lower acceptance rates in the blind sample than in the nonblind sample; in some cases, the acceptance rate dropped by more than 7 percentage points. She found similar differences with referee ratings between SBR and DBR.

A retrospective study of single-blind reviews for the *Journal of Pediatrics* and published in *JAMA* found only partial evidence for status bias, that "for the 147 brief reports, lower institutional rank was associated with lower rates of recommendation for acceptance by reviewers ($P < .001$). ... For the 258 major papers, however, there was no significant relationship between institutional rank and either the reviewer's recommendations (P=.409) or the acceptance rate (P=.508)" [Garfunkel et al. 1994, page 138].

Another retrospective analysis of single-blind reviews also published in *JAMA* found evidence of status bias at a coarse geographical level [Link 1998]. In this analysis of original research articles submitted to *Gastroenterology* during 1995 and 1996, it was found that "reviewers from the United States and outside the United States evaluate non-US papers similarly and evaluate papers submitted by US authors more favorably, with US reviewers having a significant preference for US papers" [Link 1998, page 246].

The experimental evidence is mixed concerning status bias present for top-ranked authors and institutions. The evidence is quite compelling that status bias is possible, perhaps prevalent, in SBR for most other authors and institutions, presumably for those papers most needing the critical evaluation of reviewers.

4.2 Fairness to Prolific Authors

There have been several studies that have looked at the impact of blinding on prolific authors, with conflicting results.

The Mahoney experiment discussed in the previous section also suggested that "self-citation may be a determinant of a reviewer's evaluation of a manuscript" [Mahoney et al. 1978, page 70]. In half of the papers sent to volunteer reviewers, "the author defended his contentions by referencing three of his own "in press" publications. For the other half, these same prepublication references were also cited, but were attributed to someone else. ... Reviewers rated the article as more innovative and publishable if the fictitious author included self-references in the manuscript than if no self-references were included" [ibid].

An experiment published in *JAMA* on 57 consecutive manuscripts submitted to the *Journal of Development and Behavioral Pediatrics* that were randomly assigned to either blinded or unblinded review (that is, using editorial masking) found that "contrary to the original hypothesis of this study, senior authors with more previous articles received significantly better scores from the blinded reviewers (r=−.45), but not from the nonblinded reviewers (r=−.14)" [Fisher et al. 1994, page 145]. The authors "interpret this finding to indicate that the blinded reviewers, especially those who were really blinded and could not guess author identity, may have recognized improved quality in the work of those authors with more previous publications. In contrast, reviewers who were aware of author identity did not give better scores to the more experienced authors, likely indicating that various types of bias may have entered into their thinking" [ibid, page 146].

A retrospective study of two database conferences [Madden & DeWitt 2006] found no impact on prolific authors. This study considered papers authored by those designated variously as a "famous person" or "prolific researcher" or "more senior researcher," defined as "those individuals who have published 20 or more papers in SIGMOD and VLDB conferences" [ibid, page 29]. The analysis found that DBR reviewing (with author masking) in the ACM SIGMOD conference "has had essentially no impact on the publication rates of more senior researchers in the database field" [ibid, page 30]. This result mirrors those of Blank's study, which found that acceptance rates for top-ranked institutions (where presumably most of these prolific researchers resided) were not affected by DBR.

An independent analysis of this data, using medians rather than means, reached the opposite conclusion: "that double blind review in SIGMOD do have its impact on the performance of 'famous person' " [Tung 2006]. The reason for the differing conclusions over the same data may be that this data, consisting of yearly counts of papers by prolific authors and by others, is too coarse to make a final determination, as it doesn't taken into consideration the varying submission rates of individuals and the varying participation by prolific authors.

These contradictory results render it impossible to say anything definitive about the impact of blinding on prolific authors. However, there does seem to be evidence of some kinds of bias with SBR.

4.3 Gender Equity

When reviewers know the identity of the author(s) of the submitted manuscript, gender bias is also a possibility. Several disciplines have launched in-depth studies based on concerns of gender equity.

A classic and much-referenced study showed that even when the work of a woman was identical to that of a man, the former was judged to be inferior [Goldberg 1968]. In this study, scholarly essays in a number of academic fields were presented to female college students. All of the students rated the same essays, but half of them rated essays bearing the names of male authors (e.g., John T. McKay), whereas the other half rated the same essays with the names of female authors (e.g., Joan T. McKay). The results indicated that those essays where the author was male was rated higher.

A quantitative meta-analysis of this and similar studies (over one hundred) over the intervening two decades found that "the average difference between ratings of men and women is negligible" [Swim et al. 1989, page 409]. Consistent with this analysis, 73% of studies found no significant effect for the Joan-John manipulation, 20% found that John's work was rated higher, and the remaining 7% found that Joan's work was rated higher. Interestingly, "there was some indication, however, that women will be rated less favorably than men when less information is presented" [ibid, page 421] and "there was also some indication of greater bias when the stimulus material was a resumé or application" [ibid, page 422]. More relevant to the issue of peer review is the observation from the APA task force report that when "Joan and John's work was high in quality, the effect size was close to zero (-.02 [the negative sign indicating a lower evaluation of female-authored work]); the effect was larger when Joan and John's work was medium in quality (-.24). ... these results seem to indicate that evaluation of absolutely outstanding articles will not be biased, but articles of ambiguous merit may be judged based on the author's gender" [Fouad et al. 2000, page 45]. This is again consistent with previously-discussed studies that considered status bias.

Blank's experiment, described earlier, was in fact initiated due to concerns of gender bias. The *American Economic Review* journal had employed SBR for most of its recent history, except during a period of 1973–1979 when the then-current editor adopted DBR [Borts 1974]. In the mid-1980's, the American Economic Association's Committee on the Status of Women in the Economics Profession formally expressed its concern about "the potential negative effect on women's acceptance rates of a single-blind system" [Blank 1991, page 1045]. As a result, Blank was asked by the current editor of the *AER* and the Board of Editors to design and run a randomized experiment looking into this potential effect.

Due to the careful randomization design of this experiment, one can compare acceptance rates between the blind and nonblind samples, and indeed, there were striking differences. "For women, there is no significant difference in acceptance rates between the two samples. For men, acceptance rates are significantly higher in the nonblind sample." [ibid, page 1053]. When reviewers knew that that paper was authored by a male, they accepted a higher percentage (15%, versus 11%) than if the paper was blinded. "One can compare acceptance rates between the blind and nonblind samples without other control variables because the randomization process guarantees that papers by women (and men) in each sample have identical distributions of characteristics" [ibid].

Blank emphasized the core issue: "whether the *ratio* of male to female acceptance rates in the nonblind sample is different from that in the blind sample. In both samples, women's acceptance rates are lower than men's, but the differential in the blind sample is smaller. While women in the blind sample have an acceptance rate only 1 percentage point below that of men, their rate is 3.8 percentage points lower in the non-blind sample" [ibid]. Here the results were statistically insignificant, perhaps because there were too few observations of papers authored by women.

Would DBR result in a large increase in acceptances of papers by women? "While there is some indication in these data that women do slightly better under a double-blind system, both in terms of acceptance rates and referee ratings, these effects are relatively small and statistically insignificant. Thus, this paper provides little evidence that moving to a double-blind reviewing system will substantially increase the acceptance rate for papers by female economists" [ibid, page 1063]. Interestingly, the *American Economic Review* now employs DBR.

The Modern Language Association's (MLA) experience was striking: going to DBR resulted in a large increase in acceptances by female authors. "Contributed

papers at MLA meetings had first to survive a review stage before acceptance to be read. Prior to 1974, these papers were refereed with the author's name intact. In 1974, double-blind refereeing was tried with the effect that the number of women and of new investigators having papers accepted doubled from previous years. This number doubled again when repeated in 1975, until, by 1978, the proportion of acceptances among women and new researchers was comparable to that for men. The MLA Board subsequently decided in 1979 to use double-blind refereeing for all their publications" [Billard 1993, page 321]. The impetus for this change was the perception of gender bias. "A number of women complained to the Modern Language Association in the United States that there were surprisingly few articles by women in the association's journal, compared to what would be expected from the number of women members. It was suggested that the review processes were biased. The association vigorously denied this but under pressure instituted a blind reviewing procedure under which the names of the authors and their institutional affiliations were omitted from the material sent to the reviewer. The result was unequivocal: There was a dramatic rise in the acceptance of papers by female authors" [Horrobin 1982, page 217].

It is possible that the small observed effect in Blank's study (in contrast to the MLA experience) was due to the low number of submissions by women to *AER*.

These studies show that revealing author identity, specifically the gender of the author, can sometimes have an effect on acceptance rates.

4.4 The Perception of Fairness

A *perception* of possible bias may be just as damaging as actual bias.

The Institute of Mathematical Statistics (IMS) New Researchers' Committee (NRC) report stated, "The NRC feels that the current system [SBR] has the potential for bias or perceived bias against NRs [new researchers], women and identifiable minorities, (a disproportionate number of the latter two categories are NRs)" [Altman et al. 1991, page 165]. In a response to discussants of that report, the NRC reasserted a year later, that "much of the value of double-blind refereeing lies in the community perception of fairness" [Altman et al. 1992, page 266].

The experience with this controversy at the IMS indicated a split between new researchers, which "strongly endorses double-blind refereeing. ... It seems likely that [this] represents the majority opinion among new researchers, although support for double-blind refereeing is not unanimous among new researchers" and senior members: " 'negative but sympathetic' ... seems to be a majority view among those senior enough to have been involved in the editing process" [Cox et al. 1993, page 311]. However, a survey to IMS members "indicates strong support for double-blind refereeing in the IMS journals" [ibid].

A responder to the IMS report [Cox et al. 1993] stated, "Refereeing is *perceived* by many writers as being subject to various kinds of biases: biases in favor of male or female, young or established, national or foreign researchers, working at small or large institutions, in well-developed or developing countries and so on. Whether such biases are sufficiently strong and widespread to distort the whole review process is beyond the point. So long as the *potential* for abuse is there, we should guard against it, and double-blind refereeing is but one means of ensuring such protection" [Genest 1993, page 324] (emphasis in original).

5 Quality of Reviews

Does blinding impact the quality of reviews? Two counter-balancing effects have been claimed. One possibility is that SBR, by revealing the authors' identity, increases the quality of reviews by supplying to the peer reviewer relevant information about the prior accomplishments and about publication and citation rates. On the other hand, such identity information might be used by reviewers as short cuts, thus reducing review quality. Perhaps DBR, in not revealing the author's identity, permits the reviewer to focus more on the paper itself, which can increase the quality of the review. We examine the scientific evidence of each effect in turn. Overall, the evidence is mixed on both effects.

One study [Abrams 1991] investigated a related question: whether peer evaluations of grant proposals are the best available predictor of future output of influential science. While reviewing grant proposals is different than reviewing papers submitted to refereed conferences and journals, some of these results do have relevance here.

This study examined the commonly-held perception that "individual scientists seldom fluctuate between periods of producing large amounts of good work and periods when they produce only a small amount of poor-quality work. This is the basis for hiring and promotion decisions at universities and research institutions" [ibid, pp. 112–3]. The study found a high degree of correlation. Data concerning the members of the US National Science Foundation Ecology Panel in 1988 was examined. Among the eleven senior members of this panel, high correlations between citation rates over time as well as high temporal correlations in publication rates were observed. For a group of forty-five scientists drawn from a 1973 ecology textbook who had published papers in major ecology journals, high temporal consistency among the top-ranked individuals was again observed. The conclusion is that

"scientists who have done large amounts of good quality work in the recent past are likely to continue doing so in the near future" [ibid, page 115].

There is evidence that even with this potentially useful information, journals using SBR "publish a larger fraction of papers that should not have been published than do journals" using DBR [Laband 1994, page 147]. This study used nonlinear regression and ordered probit techniques to estimate the impact of DBR on citations of a sample of 1051 articles published in 28 economics journals during 1984. The analysis found that "articles reviewed single-blind are less likely than those reviewed double-blind to be identified correctly as the highest-impact articles (those with nine or more citations in the ensuing 5 years). By the same token, articles reviewed single-blind are more likely than those reviewed double-blind to be misidentified as the lowest-impact papers (those with no citations in the ensuing 5 years). ... We conclude that the single-blind review process apparently suffers from a type I error bias to a greater extent than the double-blind review process." [ibid, page 149].

Several limitations of the Laband study have been pointed out: "There are difficulties with this analysis, the main one being that the papers considered were only reviewed in one way, either blinded or not blinded. Also, controlling for the status of the journal in which each papers appeared is inevitably a difficult process. Papers selected for the 'market leader' journals by whatever process must be more likely to be cited than those selected for more specialised or less well respected competitors" [Poutney 1996, page 1059].

The evidence is thus very mixed about whether information about prior accomplishments, coupled with the observed correlation with future accomplishments, results in better judgment about a specific submission before a reviewer. "Some argue that information about the authors' institutional affiliation helps referees evaluate manuscripts because they constitute presumptive "proof" that the research described was actually done" [Campanario 1998b, page 295]. It has also been observed that "Some referees believe that they can judge better if they know the author because the manuscript can be evaluated in the context of the author's entire corpus of work, but this claim is rare. More frequent is advocacy of anonymity for authors" [Dalton 1995, page 236]. Another asserted, "it should be the work itself, and not the reputation of the author, which influences ... As statisticians, one of our maxims is that the data should speak for themselves, so likewise should we let the work speak for itself without undue influence from outside pressures" [Billard 1993]. We now examine the scientific evidence that DBR can increase review quality.

A study of sixty articles drawn from the *Journal or Abnormal Psychology* found that "the articles by scholars affiliated with high-status institutions were cited considerably more often than the articles by scholars at low-status institutions" [Perlman 1982]. "Therefore, it appears that an institution's prestige is a valid predictor, and editors may be justified in using this as a factor in their decision making. Advocates of blind review, however, may still object to using either institutional affiliation or an individual's reputation as criteria in selecting articles. They could claim that the excellence of the manuscript should not only be apparent over time, it should also be immediately apparent without the aid of status cues. Thus, even with a blind review process, assessors should identify a higher proportion of items submitted by scholars at prestigious institutions as worthy of publication" [ibid].

Another double-blind study of DBR, carried out at the *Journal of General Internal Medicine*, found that "blinding reviewers improves the quality of review from the editor's perspective" [McNutt et al. 1990, page 1375]; see also [Evans et al. 1990]. Specifically, "editors graded the quality of blinded reviews better on three of the four quality dimensions from the editor's perspective: importance of the question, targeting key issues, and methods (all $P < .02$). The greatest difference was noted in the grades for methods. Editors graded blinded reviews from the author's point of view statistically significantly better on only one of the five quality measures: the blinded reviewer was graded as more knowledgeable (P=.05). The grades on the other dimensions of quality favored the blinded reviewer, except for courteousness. The editors' summary grades, taking into account both editor's and author's points of view, favored the blinded reviewers. The mean summary grade was 3.5 for blinded reviewers and 3.1 for unblinded reviewers. The mean difference between the blinded and unblinded reviewers was 0.41 (P=.007). The difference between the median grade for blinded and unblinded reviewers was 4.0 vs 3.0, respectively, an entire grade" [McNutt et al. 1990, pp. 1373–4]. Two limitations have been noted [Fletcher & Fletcher 1997]. "One difficulty with the study is that the referees receiving the blinded copy of the manuscript would have been aware that they were part of an experiment and may in consequence have been more careful with their reports" [Cox et al. 1993, page 316]. "The study recognized that the results may have been influenced by the nature of the journal – not a market leader, but with a very wide editorial remit – and that quite different results might be found in similar analyses of large journals, sub-specialty journals and basic science journals" [Poutney 1996, page 1059].

However, four other studies did not find that masking peer reviewers to author identity improves the quality of peer review.

In an experiment also published in *JAMA*, two randomizations were performed for submissions

over a six-month period to five biomedical journals [Justice et al. 1998]. The first assigned a quarter of the submissions to the journal's usual practice, which for four of the journals was SBR. For the rest of the submissions, one of the two reviewers was randomly selected to receive a manuscript that had been masked, by "removing author and institutional identity from the title page, running headers or footers, and acknowledgments of the manuscripts. Self-references in the text were not removed" (that is, editorial masking) [ibid, page 241]. Questionnaires were provided to editors, authors, and reviewers. Analysis of these questionnaires revealed that "authors and editors perceived no significant difference in quality between masked and unmasked reviews. We also found no difference in the degree to which the review influenced the editorial decision. ... When analysis was restricted to manuscripts that were successfully masked, review quality as assessed by editors and authors still did not differ" [ibid, page 240].

Three other studies [Tobias & Zibrin 1978, van Rooyen 1999, Smith et al. 2002], in the fields of specialized medicine and education, found similar, negative results

It is important to note that these studies utilized editorial masking. We will show shortly that masking success depends highly on how that masking is done. The authors of the *JAMA* study speculate that "poor overall masking success, in combination with the observation that an author's renown is strongly associated with masking failure," may contribute to this lack of a difference between unblinded and blinded reviews [Justice et al. 1998, page 242].

One must conclude that the jury is still out. It has not been shown convincingly that either SBR or DBR can, by revealing or by hiding the identity of the author and institution, increase the quality of the reviews of a submitted manuscript.

6 Efficacy of Blinding

Any benefits ascribed to double-blind reviewing assume that the blinding of the submitted manuscript has been successful, that reviewers cannot in fact identify the author(s) nor their institutions. "How truly anonymous any party can be in a world in which referees are selected for their in-depth knowledge of a small slice of the universe of knowledge is open to question" [Dalton 1995, page 236]. "The notion that an experimented referee can identify the author of a given paper in a specialty journal has been used by many to derogate the claim of an advantage to double-blind review" [Campanario 1998b, page 295].

We earlier differentiated editorial and author blinding. Each can be done in various ways. "To simply block out the name and affiliation from the title page requires minimal effort, to block out self-references adds a little more, and scrutinizing the manuscript for any internal cues necessitates laborious line-by-line study. Therefore, the efficacy of the blinding process will vary directly with the effort expended on it" [Pitkin 1995, page 781].

Several studies have examined how well blinding works. These studies, across a wide range of disciplines, observed that blinding achieved success rates of 53% to 79%. We now review the studies chronologically.

A study of reviewers of papers submitted to the *Journal of Social Service Research* during 1978 showed that "in 56% of the reviews, the referees did not venture a guess as to the identity of the author. In another 4%, the referees guessed wrong. In an additional 5%, the referees made correct guesses about some bit of identifying information, but they did not guess the name of the author" [Rosenblatt & Kirk 1980, page 389]. This works out to successful blinding 65% of the time, for editorial blinding in which the names of authors and their institutional affiliations are removed from the title page.

One retrospective study over six journals employing DBR with author blinding and representing a broad range of areas in psychology showed that "35.6% of the 146 reviewers were correct in their identifications of the author or of at least one of the authors in the case of multiauthored papers. ... There were no significant differences in the proportion of correct detections among the six journals, ranging from 26% to 42% ... nor was there any relationship between detection accuracy and the number of years of reviewing experience" [Ceci & Peters 1984, page 1493]. When "editorial staff oversights in not removing title pages of manuscripts before sending them to reviewers or authors' oversights in preparing their manuscripts, such as explicit flagging of former work ("In our earlier work ...") or inappropriate inclusion of personal acknowledgments in the body of the text ... are excluded from the analysis, overall only 25.7% of reviewers are able to detect authors' identities, with very little variation among the six journals" [ibid].

In the McNutt study described in the previous section, editorial blinding was successful to institution name for 73% of the reviewers and to author(s)' names for 76% of thr reviewers [McNutt et al. 1990]. The blinding process was of moderate thoroughness: "An editorial assistant copied each manuscript, retyped the title page, and removed authors' and institutions' identifiers using an opaque tape. To do this, the entire manuscript was scanned—headers and footers, body of text, tables, and figures. She made no attempt to remove references to the author's own work. ... Minimal changes were made, on average, to the body of the manuscript" [ibid, page 1372].

Submissions to the *American Journal of Public Health (AJPH)* are partially author blinded and partially editori-

ally blinded: "Contributors to *AJPH* are instructed to submit a second face sheet which includes only the title of the paper. These instructions are usually followed. We remove acknowledgments, but make no further effort to remove identifying page headers (when present contrary to instructions) or to change the text or references. We have been aware that a substantial portion of our manuscripts are not truly blinded because of text allusions and self-referencing" [Yankauer 1991, page 843–4]. A questionnaire revealed that "blinding could be considered successful 53% or 61% of the time, depending on whether successful blinding ignores identification or includes only correct identification. ... Self-referencing (61.8%) and personal knowledge (38.2%) were the two clues given for identification of author and/or institution. In both cases 16% of the indentifications were incorrect" [ibid, page 844].

Blank's experiment found that "among all referee surveys received for blind papers, slightly over half (50.9 percent) claim to know the author. Ten percent of these referees are incorrect, however, so that only 45.6 percent of the authors in the blind sample are correctly identified. (Multiple-author papers are considered to be correctly identified in any of the author's names are known.)" [Blank 1991, page 1051] Note that in this experiment the papers were editorially blinded only by changing the first or second pages.

The study by Fisher et al. discussed in the previous section also considered (editorial) masking success, "in which the cover page and any identifying data on the top of bottom of each page had been removed; so as not to alter the quality of the manuscript, no effort was made to delete information in the text when the authors might have identified themselves" [Fisher et al. 1994, page 144]. 54% of reviewers were thereby successfully blinded.

The study by Justice et al. also discussed in the previous section found that with (editorial) masking, "manuscripts by authors with whom the unmasked reviewer was familiar ... were less likely to be successfully masked (53%) (that is, the masked reviewer was more likely to correctly guess author identity) than those of authors who were not known to the unmasked reviewer (79%)" [Justice et al. 1998, page 241].

Another experiment in *JAMA* found that "a long-standing policy of masking did not increase masking success" [Cho et al. 1998, page 243]. This study included four medical journals that did not mask author identity and three medical journals with a policy of DBR. Papers were then editorially blinded: "Each journal masked eligible manuscripts by removing author and institutional identity from the title page, running headers or footers, and acknowledgments of manuscripts. Self-references in the text were not removed" [ibid, page 244]. 60% of reviewers were masked. "There was no significant difference in masking success between journals with a policy of masking and those without (P=.92)" [ibid, page 245].

Another randomized study from a few years ago in a medical journal showed that "with successful blinding defined as either author not identified or author identified incorrectly, 170 reviewers (58%) were successfully blinded" [van Rooyen 1999, page 235]. Blinding was editorial: "Blinding consisted of removing authors' details from the title page and acknowledgments. No attempt was made to remove authors' details from within the text of the manuscript, the illustrations, or the references" [ibid, page 234].

Most of these studies utilized editorial masking, achieving success rates of 53% to 79%, with a gross average across studies of 62%. The success rate was lower for known authors. Self-referencing was a major clue to reviewers. And there were incorrect guesses. So even minimal editorial blinding can be somewhat effective. "Clearly, the feasibility and success of blinding depends both on the amount of effort put into the blinding process and on factors related to the type and circulation of the involved journal" [Fisher et al. 1994, page 145].

A data mining experiment took a different tack, seeing whether a computer program could identify authors using only the citations included in the paper. Two automatic methods for author identification were considered: "(1) a (dynamic) vector-space model that represents both papers and author histories, and (2) tallying (discriminative) self-citations." [Hill & Provost 2003, page 179]. A very large archive of physics papers gathered as part of the KDD Cup 2003 competition was used in the study. "The self-citation based methods generally worked better. However, the vector-space models are able to match (with much lower accuracy) even when self-citations are removed. With the best method, based on discriminative self-citations, authors can be identified 45% of the time. Additionally, the top-10% most prolific authors can be identified 60% of the time. ... authors with 100 or more prior publications can be identified 85% of the time" [ibid].

Blank's conclusions apply to the many studies generally. "On the one hand, a substantial fraction—almost half—of the blind papers in this experiment could be identified by the referee. This indicates the extent to which no reviewing system can ever be fully anonymous. On the other hand, more than half of the papers in the blind sample *were* completely anonymous. A substantial fraction of submitted papers are not readily identified by reviewers in the field. ... Those blind papers that are correctly identified by the referees ... are skewed in favor of authors who are better known or who belong to networks that distribute their working papers more widely" [Blank 1991, pp. 1051–2]. Ceci and Peters conclude that "Although there are occasional lapses in the preparation of manuscripts by au-

thors and failures to screen manuscripts by editorial staff, we are impressed by the overall efficiency of blind review" [Ceci & Peters 1984, page 1494].

7 Recommendations of Scholarly Societies

Several scholarly societies have weighed in on this question.

An examination published in the *Journal of Business Ethics* of past abuses of the editorial process warranted the proposal to "use the double blind process in which referees do not know authors nor authors, referees" so as to "protect referees from influence and participants from damage in a negative decision" [Carland et al. 1992, page 103].

The Institute of Mathematical Statistics (IMS) New Researchers' Committee (NRC) report stated, "after extensive discussion, the consensus of the NRC is that the advantages of the double-blind system outweigh the costs, and we recommend that IMS journals evaluate the benefits of adopting such a system" [Altman et al. 1991, page 166]. In a response to discussants of this report, the NRC reasserted a year later, that "the NRC strongly supports double-blind refereeing for its potential to remove separate consideration, perceived or otherwise" [Altman et al. 1992, page 266].

The view of the IMS Ad Hoc Committee on Double-Blind Refereeing in 1993 "may be summarized as cautiously receptive to double-blind refereeing. We are not convinced that the benefits outweigh the disadvantages; but we are open to the possibility. We recommend that if a change in journal policy is contemplated that an experiment be conducted to assess the merits of double-blind refereeing before any permanent change is made" [Cox et al. 1993, page 311].

The APA Task Force on Women in Academe states that "Because of the potential for bias, APA has mandated that editors of APA journals offer masked review as an option; however, mandatory masked review of articles should be instituted as policy" [Fouad et al. 2000, page 34]. The report includes in its recommendations, to "Implement a policy of mandatory masked review for all APA peer-reviewed publications" [ibid, page 45]. The APA Guide states, "Most APA journals, like the majority of other professional publications, practice anonymous, or masked reviews" [Calfee & Valencia 2006]. Of the 47 APA journals that accept paper submissions, almost half require masking and most of the rest allow it on request.

There is a pattern here. "It is therefore only to be expected that senior established researchers will tend to seek the status quo, being less inclined to want to move to double-blind refereeing, while new (and also women and researchers in lower status named institutions) researchers will tend to prefer that double-blind refereeing be introduced" [Billard 1993, page 322].

8 Prevalence

Disciplines vary widely in their use of single- and double-blind review, but the historical trend is clear.

A non-randomized survey in 1989 showed that "in chemistry, physics, math, and psychology, the responding journals indicate that they use a single-blind reviewing system. ... Biology appears to have both single-blind and double-blind journals, as does history and anthropology. Political-science and sociology journals report uniformly using double-blind reviewing methods" [Blank 1991, pp. 1043–45]. Overall, considering 37 journals in nine disciplines, 79% were single-blind and 21% were double-blind.

"Four surveys of the frequency of binding have been published, but none is based on a random sampling of journals, and their size and response rates often leave something to be desired. Their results suggest that the majority of scientific journals do not practice blind review and that blinding may be more common in the social sciences than in the physical and medical sciences" [Yankauer 1991, page 843].

A more recent survey of 553 journals selected from eighteen disciplines revealed that DBR is increasing in prevalence, in comparison with previous studies: "Across the disciplines, the majority of surveyed journals used double-blind reviews (58%), 37% employed single-blind, and only 5% made use of open review" [Bachand & Sawallis 2003, page 54]. Within computer science (29 journals, 15 responding), 57% were single-blind and 43% were double-blind.

Within ACM, all journals are single-blind and about 80% of ACM conferences are single-blind (as of 2000 [Snodgrass 2000]).

Similarly, the database field has traditionally relied on single-blind reviewing. Until recently, all of its conferences and journals have been single-blind. The ACM SIGMOD conference adopted double-blind reviewing in 2001 [ibid].

9 Summary

"There is a long tradition attached to the peer review system. As *users* of science, we all depend on it: our professional realizations are based upon the work of others, and we count on journal (and book) editors to separate the wheat from the tares. Although there is no such thing as perfection, it would be a disservice to the profession if too many scientific writings addressed irrelevant issues or

contained gross factual errors. As *producers* of science, it is also in our interest that the system be fair: favoritism, discrimination and condescension bring discredit on the entire operation and ultimately work against the discipline, even if individual benefits occasionally may accrue in the short term" [Genest 1993, page 324].

As the noted statistician Lynne Billard, who has written extensively on this topic, has remarked, "The issue of double-blind refereeing today is one fraught with emotional overtones both rational and irrational, often subconsciously culturally based, and so is difficult for many of us to resolve equitably no matter how well intentioned" [Billard 1993, page 320].

We have attempted here to summarize the many studies of the varied aspects of blind reviewing within a large number of disciplines.

Concerning the central issue of fairness, Blank's summary in 1991 of the literature still holds true fifteen years later. "In summary, the literature on single-blind versus double-blind reviewing spans a wide variety of disciplines and provides rather mixed results. Few of the empirical tabulations provide convincing evidence on the effects or non-effects of refereeing practices, largely because of their inability to control for other factors in the data. If not fully convincing, however, there is at least a disturbing amount of evidence in these studies that is consistent with the hypothesis of referee bias in single-blind reviewing" [page 1045]. Many studies provide evidence that DBR is fairer to authors from less-prestigious institutions and to women authors. Such differences are likely to matter even more for highly-selective conferences and journals.

Concerning quality of reviews, it is not known definitely whether either SBR or DBR results in a higher quality of reviews.

Most of the studies discussed here utilize editorial blinding, which has been shown to be successful about 60% of the time, across many disciplines. Removing text allusions and self-citations would increase success rates to perhaps 75%.

The prevalence of DBR has increased dramatically over the last fifteen years, to the point where most scientific journals now employ double-blind reviewing.

There is an administrative cost to DBR, to the journal as well as to the author. These costs vary depending on how the blinding is done, with the efficacy directly related to the effort expended.

Journals strongly desire to fairly evaluate submitted manuscripts, while simultaneously keeping costs in control. The policy question before each journal and each scholarly publisher is thus the following. Is the documented benefit of equity worth the administrative cost? At what price fairness?

Acknowledgments

The author thanks Tamer Özsu for comments on a previous version and Merrie Brucks for her help throughout.

References

[Abrams 1991] P. A. Abrams, "The Predictive Ability of Peer Review of Grant Proposals: The Case of Ecology and the US National Science Foundation," *Social Studies of Science* 21(1):111–132, February 1991.

[Altman et al. 1991] N. Altman, D. Banks, P. Chen, D. Duffy, J. Hardwick, C. Léger, A. Owen, and T. Stukel, "Meeting the Needs of New Statistical Researchers," *Statistical Science* 6(2):163–174, May 1991.

[Altman et al. 1992] N. Altman, J. F. Angers, D. Banks, D. Duffy, J. Hardwick, C. Léger, M. Martin, D. Nolan (Chair), A. Owen, D. Politis, K. Roeder, T. N. Stukle and Z. Ying, "Rejoinder," *Statistical Science* 7(2):265–266, May 1992.

[Armstrong 1997] J. S. Armstrong, "Peer Review for Journals: Evidence on Quality Control, Fairness, and Innovation," *Science and Engineering Ethics* 3:63–84, 1997.

[Bachand & Sawallis 2003] R. G. Bachand and P. P. Sawallis, "Accuracy in the Identification of Scholarly and Peer-Reviewed Journals and the Peer-Review Process Across Disciplines," *The Serials Librarian* 45(2):39–59, 2003.

[Billard 1993] L. Billard, "Comment," *Statistical Science* 8(3):320–322, August 1993.

[Calfee & Valencia 2006] R. C. Calfee and R. R. Valencia, "APA Guide to Preparing Manuscripts for Journal Publication," `http://www.apa.org/journals/authors/guide.html`, viewed June 21, 2006.

[Blank 1991] R. M. Blank, "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review," *American Economic Review* 81(5):1041–1067, December 1991.

[Borts 1974] G. Borts, "Report of the Managing Editor," *American Economic Review* 64:476–82, May 1974.

[Campanario 1998a] J. M. Campanario, "Peer Review for Journals at It Stands Today—Part 1," *Science Communication* 19(3):181–211, March 1998.

[Campanario 1998b] J. M. Campanario, "Peer Review for Journals at It Stands Today—Part 2," *Science Communication* 19(4):277–306, June 1998.

[Carland et al. 1992] J. A. Carland, J. W. Carland, and C. D. Aby, Jr., "Proposed Codification of Ethicacy in the Publication Process," *Journal of Business Ethics* 11:95–104, 1992.

[Ceci & Peters 1984] S. J. Ceci and D. P. Peters, "How blind is blind review?" *American Psychologist* 39:1491–1494, 1984.

[Cho et al. 1998] M. K. Cho, A. C. Justice, M. A. Winker, J. A. Berlin, J. F. Waecklerle, M. L. Callaham, and D. Rennie, "Masking author identity in peer review—What factors influence masking success?" *Journal of the American Medical Association* 280(3):243–245, July 15, 1998.

[Cox et al. 1993] D. Cox, L. Gleser, M. Perlman, N. Reid, and K. Roeder, "Report of the Ad Hoc Committee of Double-Blind Refereeing," *Statistical Science* 8(3):310–317, August 1993.

[Dalton 1995] M. Dalton, "Refereeing of Scholarly Works for Primary Publishing," in **Annual Review of Information Science and Technology (ARIST)**, Volume 30, M. E. Williams (ed.), American Society of Information Science, 1995.

[Evans et al. 1990] A. T. Evans, R. A. McNutt, R. H. Fletcher, and S. W. Fletcher, "The Effects of Blinding on the Quality of Peer Review: A Randomized Trial," *Clinical Research* 38(2), 1990.

[Fisher et al. 1994] M. Fisher, S. B. Friedman, and B. Strauss, "The effects of blinding on acceptance of research papers by peer review," *Journal of the American Medical Association* 272(2):143–146, July 13, 1994.

[Fletcher & Fletcher 1997] R. H. Fletcher and S. W. Fletcher, "Evidence for the Effectiveness of Peer Review," *Science and Engineering Ethics* 3(1):35–50, 1997.

[Fouad et al. 2000] N. Fouad, S. Brehm, C. I. Hall, M. E. Kite, J. S. Hyde, and N. F. Russo, **Women in Academe: Two Steps Forward, One Step Back**, Report of the Task Force on Women in Academe, American Psychological Association, 2000. `http://www.apa.org/pi/wpo/academe/repthome.html`, viewed June 21, 2006.

[Franzini 1987] L. R. Franzini, "Editors Are Not Blind," *American Psychologist*, page 104, January 1987.

[Garfunkel et al. 1994] J. M. Garfunkel, M. H. Ulshen, H. J. Hamrick, and E. E. Lawon, "Effect of Institutional Prestige on Reviewers' Recommendations and Editorial Decisions," *Journal of the American Medical Association* 272(2):137–138, July 13, 1994.

[Genest 1993] C. Genest, "Comment," *Statistical Science* 8(3):323–327, August, 1993.

[Goldberg 1968] P. Goldberg, "Are some women prejudiced against women?" *Transaction* 5:28–30, April, 1968.

[Gordon 1980] M. D. Gordon, "The role of referees in scientific communication," in **The Psychology of Written Communication: Selected Readings**, J. Hartley (ed.), London, England: Kogan Page, pp. 263–275, 1980.

[Harnad 1982] S. Harnad, "Peer commentary on peer review," *Behavioral and Brain Sciences* 5(2):185–186, 1982.

[Hill & Provost 2003] S. Hill and F. Provost, "The Myth of the Double-Blind Review," *SIGKDD Explorations* 2(5):179–184, December 2003.

[Horrobin 1982] D. F. Horrobin, "Peer review: A philosophically faulty concept which is proving disastrous to science," *Behavioral and Brain Sciences* 5(2):217–218, 1982.

[Justice et al. 1998] A. C. Justice, M. K. Cho, M. A. Winker, J. A. Berlin, and D. Rennie, "Does Masking Author Identity Improve Peer Review Quality?" *Journal of the American Medical Association* 280(3):240–242, July 15, 1998.

[Laband 1994] D. N. Laband, "A Citation Analysis of the Impact of Blinded Peer Review," *Journal of the American Medical Association* 272(2):147–149, July 13, 1994.

[Link 1998] A. M. Link, "US and Non-US Submission: An Analysis of Review Bias," *Journal of the American Medical Association* 280(3):246–247, July 15, 1998.

[Madden & DeWitt 2006] S. Madden and D. DeWitt, "Impact of Double-Blind Reviewing on SIGMOD Publication Rates," *ACM SIGMOD Record* 35(2):29–32, June 2006.

[Mahoney et al. 1978] M. J. Mahoney, A. E. Kazdin, and M. Kenigsberg, "Getting Published," *Cognitive Therapy and Research* 2(1):69–70, 1978.

[McGiffert 1988] M. McGiffert, "Is Justice Blind? An Inquiry into Peer Review," *Scholarly Publishing* 20(1):43–48, October 1988.

[McNutt et al. 1990] R. A. MuNutt, A. T. Evans, R. H. Fletcher, and S. W. Fletcher, "The Effects of Blinding on the Quality of Peer Review," *Journal of the American Medical Association* 263(10):1371–1376, March 9, 1990.

[Perlman 1982] D. Perlman, "Reviewer "bias": Do Peters and Ceci protest too much?" *Behavioral and Brain Sciences* 5(2):231–232, 1982.

[Peters & Ceci 1982] D. P. Peters and S. J. Ceci, "Peer-Review Practices of Psychological Journals: The Fate of Published Articles, Submitted Again," *Behavioral and Brain Sciences* 5(2):187–195, 1982.

[Pitkin 1995] R. M. Pitkin, "Blinded Manuscript Review: An Idea Whose Time Has Come?" Editorial, *Obstetrics and Gynecology* 85(5 Part 1):781–782, May 1995.

[Poutney 1996] M. Poutney, "Blinded Reviewing," Editorial, *Developmental Medicine and Child Neurology* 38(12):1059–1060, December 1996.

[Rennie 1990] D. Rennie, "Editorial Peer Review in Biomedical Publication: The First International Conference," *Journal of the American Medical Association* 263(10):1317, March 9, 1990.

[Rennie & Flanagin 1994] D. Rennie and A. Flanagin, "The Second International Congress on Peer Review in Biomedical Publication," *Journal of the American Medical Association* 272(2):91, July 13, 1994.

[Rosenblatt & Kirk 1980] A. Rosenblatt and S. A. Kirk, "Recognition of Authors in Blind Review of Manuscripts," *Journal of Social Service Research* 3(4):383–394, Summer 1980.

[Rosenthal 1982] R. Rosenthal, "Reliability and Bias in Peer-Review Practices," *Behavioral and Brain Sciences* 5(2):235–236, 1991.

[Smith et al. 2002] J. A. Smith, R. Nixon, A. J. Bueschen, D. D. Venable, and H. H. Henry, "The Impact of Blinded versus Unblinded Abstract Review on Scientific Program Content," *Journal of Urology* 168(5):2123–2125, November 2002.

[Snodgrass 2000] R. T. Snodgrass, "Chair's Message," *ACM SIGMOD Record* 29(1):3, March 2000.

[Snodgrass 2003] R. T. Snodgrass, "Developments at *TODS*," *ACM SIGMOD Record* 32(4):14–15, December 2003.

[Swim et al. 1989] J. K. Swim, E. Borgida, G. Maruyama, and D. G. Myers, "Joan McKay versus John McKay: Do gender stereotypes bias evaluations?" *Psychological Bulletin* 105(3):409–429, 1989.

[Tobias & Zibrin 1978] S. Tobias and M. Zibrin, "Does Blind Reviewing Make a Difference?" *Educational Researcher* 7(1):14–16, January 1978.

[Tung 2006] A. K. H. Tung, "Impact of Double Blind Reviewing on SIGMOD Publication: A More Detail Analysis," 2 pages, July 2006.

[van Rooyen 1999] S. van Rooyen, F. Godlee, S. Evans, R. Smith, and N. Black, "Effect of Blinding and Unmasking on the Quality of Peer Review: A Randomized Trial," *Journal of General Internal Medicine* 14(10):622–624, October 1999.

[Yankauer 1991] A. Yankauer, "How Blind is Blind Review?" *American Journal of Public Health* 81(7):843–845, July 1991.

[Zuckerman & Merton 1971] H. Zuckerman and R. K. Merton, "Patterns of Evaluation in Science: Institutionalization, Structure and Functions of the Referee System," *Minerva* 9(1):66–100, January 1971.

Model Driven Development of Secure XML Databases

Belén Vela[1], Eduardo Fernández-Medina[2], Esperanza Marcos[1] and Mario Piattini[2]

(1) Kybele Research Group. Languages and Computing Systems Department
Rey Juan Carlos University. C/ Tulipán, s/n - 28933 Móstoles, Madrid, Spain
{belen.vela, esperanza.marcos}@urjc.es
(2) Alarcos Research Group. Information Systems and Technologies Department
UCLM-Soluziona Research and Development Institute
University of Castilla-La Mancha. Paseo de la Universidad, 4 – 13071 Ciudad Real, Spain
{Eduardo.FdezMedina, Mario.Piattini}@uclm.es

ABSTRACT

In this paper, we propose a methodological approach for the model driven development of secure XML databases (DB). This proposal is within the framework of MIDAS, a model driven methodology for the development of Web Information Systems based on the Model Driven Architecture (MDA) proposed by the Object Management Group (OMG) [20]. The XML DB development process in MIDAS proposes using the data conceptual model as a Platform Independent Model (PIM) and the XML Schema model as a Platform Specific Model (PSM), with both of these represented in UML. In this work, such models will be modified, so as to be able to add security aspects if the stored information is considered as critical. On the one hand, the use of a UML extension to incorporate security aspects at the conceptual level of secure DB development (PIM) is proposed; on the other, the previously-defined XML schema profile will be modified, the purpose being to incorporate security aspects at the logical level of the secure XML DB development (PSM). In addition to all this, the semi-automatic mappings from PIM to PSM for secure XML DB will be defined.

1 Introduction

Though relational database (DB) technology still plays a central role in the data management arena today, we have seen numerous evolutions of this technology, such as the XML DBs. A key requirement underlying those recent data management systems is a demand for adequate security. Fine-grained flexible authorization models and access control mechanisms, in particular, are being called for [1]. Traditionally, the information of XML documents was stored directly in XML files or in conventional Database Management Systems (DBMSs), by mapping the XML data to relational data stored in relational tables or by using the data types supplied for supporting file management, as for example the CLOB (Character Large OBject) type. The XML DBs are now emerging as the best alternative for storing and managing XML documents. At present, there are different solutions to store XML documents, and they could be roughly categorized, according to [25], into two main groups: native XML DBMSs like Tamino [23]; and XML DB extensions enabling the storage of XML documents within conventional, usually relational or Object-Relational (OR) DBMSs such as Oracle. This latter includes, since version 9*i* release 2, new features for the storage of XML (Oracle's XML DB) [22]. In [25] a study of different XML DB solutions is performed.

For most organizations, management, security and confidentiality of information are critical topics [6]. Moreover, as some authors remark, information security is a serious requirement which must be carefully considered, not as an isolated aspect, but as an element that is present in all stages of the development life cycle [5,11,13]. A body as important as the Information Systems Audit and Control Foundation insists on the fact that security should be considered explicitly and as an integral item in all the development stages of an information system [15]. In the case of the XML DBs, security is also a key aspect that must be explicitly considered. It has to be taken into account in an orthogonal way for the complete development process of this kind of DB. Access control models have been widely investigated and several access control systems, specifically tailored to XML documents, have been developed [2,3,4,12,14,18]. However, all of them define security criteria directly over the XML documents or DTDs.

Our approach is based on the Model Driven Architecture (MDA) proposed by the Object Management Group (OMG) and allows us to define the security specifications on the conceptual data model, independently of the target logical data model (DB schema). Starting from this secure conceptual data model we transform it semi-automatically into a secure XML DB, as a logical data model.

Although there are different ideas for integrating security into the information systems development process, information security within the scope of DBs tends to be considered only from a cryptographic point of view. Recently, we have proposed a methodology for relational DB which integrates security aspects at

all stages of the development process [7]. However, to the best of our knowledge, there are no works that deal with security when developing an XML DB.

In this paper, we will integrate the security aspect into the methodological approach for XML DB development [24] framed in MIDAS [16], a model driven methodology for the development of Web Information Systems (WIS). MIDAS proposes the use of standards in the development process, as well as the use of UML in modelling the WIS, irrespective of the abstraction level and the aspect of the system to be modelled. As UML does not allow us to represent all the necessary models, MIDAS incorporates some existing UML extensions and defines or adapts some new ones, whenever necessary [8,17].

In the next section, we will introduce the secure XML DB development process in the framework of MIDAS, where the Platform Independent Model (PIM) is the conceptual data model. It will be represented with an extended UML class diagram that includes the security aspect at this level. This profile will be summed up in section 3. As data Platform Specific Model (PSM) in MIDAS, it is proposed to use the OR model or the XML Schema model, depending on the technology used. In this paper we will show the part corresponding to secure XML DB development. The PSM employed will therefore be the XML Schema model. In section 4, we will present an adaptation of the previously-defined profile for XML DBs for the incorporation of specific security aspects into this kind of DBs. In section 5, we will show the mappings from the secure data PIM to the secure data PSM which will be the schema of the secure XML DB. These mappings are based on those defined in [24], where the rules to obtain the data PSM are described, but without taking into consideration security aspects. In this paper, we will adapt such rules so as to obtain the schema of an XML DB which includes the necessary constraints for security. Finally, in section 0, we will put forward our main conclusions and present our future work.

2 Secure XML DB Development Process

MIDAS proposes a model driven architecture based on MDA and, when modelling the system, considers, the aspects of *content*, *hypertext* and *behaviour* at the levels of Computation Independent Models (CIMs), common to all the system, PIMs and PSMs. In Figure 1 we can see the simplified MIDAS MDA.

In this paper, we will focus on the **content** aspect, which corresponds to the traditional concept of a DB, for the ***PIM*** and ***PSM*** levels. The development of a DB depends on several aspects; on the one hand, on whether there is already a DB within the organization or not, and, on the other hand, on the technology to be used: in other words, if we aim to use an OR DB [17] or an XML DB [24].

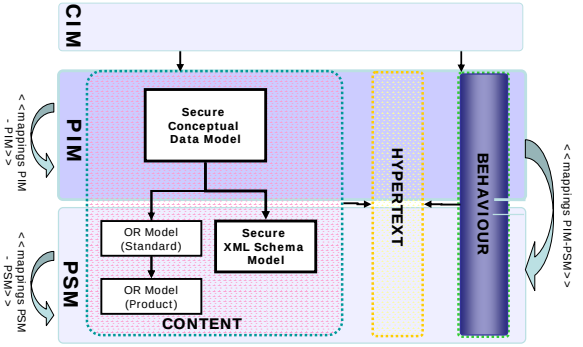

Figure 1. Simplified MIDAS Architecture

Moreover, a third dimension is considered in MIDAS, and it includes all aspects to be taken into account when developing a WIS, such as the system architecture or security. This third dimension is orthogonal to the ones presented in Figure 1.

In the cases in which the DB that we want to develop includes information to be protected, the security aspect will have to be taken into account from the first stages of the DB development. So, for the model driven development of a secure XML DB we have to perform the following tasks:

- At the ***PIM*** level, the secure data conceptual model is carried out without considering the selected technology, since this model is platform independent. This secure data PIM is represented through an extended UML class diagram, so as to be able to represent *security* aspects together with a set of security constraints that have been expressed through OSCL language [8], as we will see in the next section.
- At the ***PSM*** level, the data logical design is performed, taking into account the selected technology. In our case, this is an XML DB. We will start from the secure data PIM obtained at the previous level and will apply the mappings summarized in section 5. The secure data PSM will be represented through an XML schema in extended UML (see section 4). In this case, the DB schema will be the XML schema, which takes into account the necessary security aspects.

3 Secure Data PIM

To develop a secure data PIM, a secure UML profile has been developed (for more details, see [8]). The defined UML profile allows us to classify both data and users according to different classification criteria. These criteria are the following ones:

- **Security levels**: to define a hierarchy of levels such as those traditionally employed in the army: unclassified, confidential, secret and top secret.
- **User roles**: to define a hierarchical set of user roles that represents the hierarchical functions within an enterprise.
- **User categories**: to define a horizontal organization or classification (non hierarchical) of user groups.

In addition to this classification information, the profile allows us to define three kinds of constraints:

- **Data dynamic classification rules**: to define the classification data of different instances, depending on the value of one or several attributes of the instances.
- **Audit rules**: They specify situations in which it is interesting to us to register an audit trace to analyze which users have accessed (or have tried to access) information. To do so, conditions expressed in OCL are defined.
- **Authorization rules**: to define which users will be allowed to access to which data and to perform which actions depending on a condition expressed in OCL.

Our security model is general, and the classification criteria, together with the *data dynamic classification rules* and the authorization rules, allow us to integrate several access control models, such as the mandatory access control, a simplified role based access control, discretionary access control and access control based on rules. The coexistence of these rules frequently provokes conflicts, that we solve by applying a set of conflict resolution rules defined in [8, 9, 10].

For the definition of all these elements, we consider the UML profile known as *Conceptual Secure DB* (extension of UML and OCL to design secure DBs), which is composed of a set of data types, tagged values and stereotypes, together with the definition of a set of well-formedness rules. The package containing all the stereotypes defined within this UML profile can be analyzed in Figure 2. These stereotypes can be classified into three categories:

- The stereotypes necessary for representing security information in the *model elements.*
- The stereotypes needed to model the *security constraints* when defining: a) the dynamic classification of any element, b) audit rules expressed in OCL and c) authorization rules.
- The *UserProfile* stereotype that is necessary to specify security constraints on what might be seen as a property of a user or a group of users, for instance; citizenship, age, etc.

A detailed description of all these stereotypes, as well as the tagged values that have been defined for them, can be found in [8].

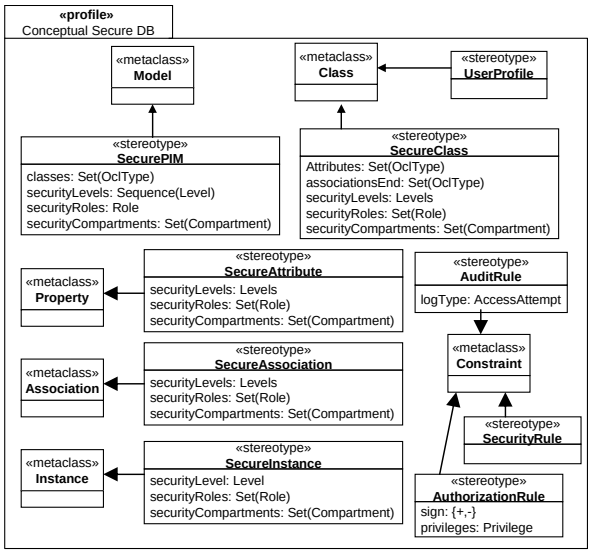

Figure 2. *Conceptual Secure DB* profile

4 Secure Data PSM

In MIDAS, the XML schema model is proposed as data PSM. It is represented in extended UML, using the profile defined in [24]. To include the security aspects in the model, in this paper we have adapted such a profile by adding the elements that are needed to be able to consider the aspect of *security*.

In Figure 3, we will show the elements that have been added, with the goal of adapting the profile so that it is able to represent secure XML schemas through a UML class diagram. The extension defines a set of new stereotypes. The aim is for it to be able to consider all the components of a secure XML in a graphical notation of UML, maintaining the associations, the order and the links between the different elements.

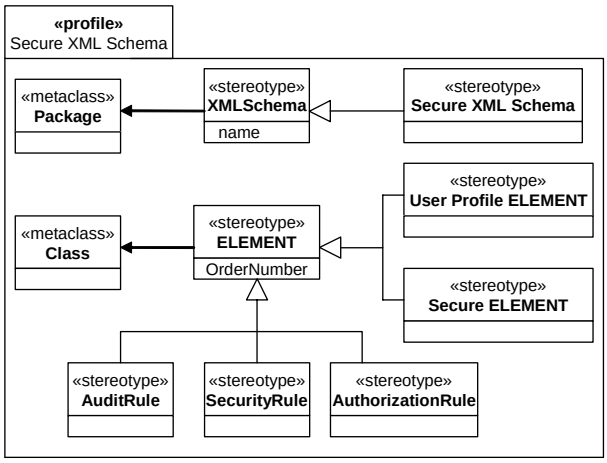

Figure 3. *Secure XML Schema* profile

5 Mappings from PIM to PSM

In the same way that methodologies for relational or OR DBs propose some rules for the transformation of a conceptual schema into a standard logical one, in MIDAS, mappings from the data PIM to the data PSM are proposed. In this work, we have defined the transformation rules needed to obtain a secure data PSM from the secure data PIM. The work of [24] is taken as a basis, where the different mappings to obtain the schema of an XML DB were defined (but where security was not yet taken into account).

- **Transformation of the secure data PIM:** The data conceptual model, that is, the *secure PIM*, is transformed, at the PSM level, into an XML schema named '*Secure Data PSM*'. It will be represented with a UML package stereotyped with <<Secure XML SCHEMA>> and will be called as the XML schema. It will include all components of the secure XML schema (PSM). Furthermore, it will contain the security attributes (*securityLevel*, *securityRoles* and *securityCompartments*) of the secure PIM. These attributes will be defined within the XML schema as global elements. They could have been included as schema attributes but if they were represented in such a way, they would not be considered first order elements and the fact that they could have a multiple maximum cardinality could not be collected either.
- **Transformation of the *User Profile* class:** This class includes the information that we want to record for each user. It will be transformed by including a global element stereotyped with <<User Profile ELEMENT>>, which will contain a sequence complexType with all class attributes as subelements.
- **Transformation of secure classes:** In a generic way, a UML class is transformed into an element of the XML schema with the same name as the class it comes from [24]. To transform secure UML classes, stereotyped with <<SecureClass>>, we have to include the secure characteristics that they have, too. Secure classes can have three specific attributes: *securityLevel*, *securityRoles* and *SecurityCompartments*. They will be transformed into secure elements stereotyped with <<Secure ELEMENT>>. Each secure element will contain a complexType of sequence type, which will contain as subelements, among others, the secure attributes, indicating, with the subelements attribute *maxOccurs*, the number of possible instances of the security attributes.
- **Transformation of secure attributes:** Due to the fact that the attributes of a class, according to the proposal of [24], are transformed as subelements of the element that represents the UML class to which those attributes belong, if an attribute has its own security attributes associated with it, these attributes will be represented as subelements of the element that represents the corresponding attribute. Thus, the security attributes defined within an attribute will be transformed into <<Secure ELEMENT>> subelements.
- **Transformation of secure associations:** Regarding the transformation of associations, a detailed study of the most appropriate way to map them at the PSM level was carried out in [24]. The associations between two classes are transformed, in a generic way, by including a subelement in one of the elements, corresponding to one of the classes implied in the relationship with one or several references to the other element implicated in the association. If it were a secure association, this subelement would have subelements to represent the corresponding security attributes (*securityLevel*, *securityRoles*, *securityCompartment*) stereotyped as <<Secure ELEMENT>>.
- **Transformation of security constraints:** When transforming the security constraints that had been defined at the PIM level, these can be defined for any element (model or class), although it is normal to define them at the class level. If they are defined at the model level, global elements to collect this fact will be created. In the rest of the cases, subelements of the elements they depend on will be created. There are three types of constraints:

a) **Audit Rules**: They will be transformed by creating a subelement stereotyped with <<AuditRule>> with the name of "AuditRule_" plus the number of the rule. This element will be of the complexType and it will contain a sequence formed by two elements: One *AuditRuleType* element of simple Type of the *string* base type with a constraint of enumeration type with the values *all*, *frustratedAttempt*, *successfullAccess*; and another element *AuditRuleCondition* that will be an element of *string* type, that will contain the XPATH expression associated with the OCL expression.

```
<complexType>
 <sequence>
  <element name= "AuditRuleType">
   <simpleType>
        <restriction base= "string">
          <enumeration value= "all"/>
          <enumeration value= "frustatedAttempt"/>
         <enumeration value= "successfullAccess"/>
        </restriction>
   </simpleType>
  </element>
  <element name= "AuditRuleCondition" type="string"/>
 </sequence>
</complexType>
```

b) **Security Rule:** The dynamic classification of any PIM element will be transformed by creating a subelement stereotyped with <SecurityRule>>, with the name "SecurityRule_" plus the number of the rule. This element will be of complexType and it will contain one element of string type with the XPATH expression associated with the OCL expression.

```
<complexType>
  <sequence>
    <element name= "SecurityRuleCondition" type="string"/>
  </sequence>
</complexType>
```

c) **Authorization Rules:** These will be transformed by creating a subelement stereotyped with <<AuthorizationRule>> with the name "AuthorizationRule_", plus the number of the rule. This element will be of a complexType and it will contain a sequence formed by three elements: An *AuthorizationRuleSign* element of simpleType of *string* base type with a constraint of enumeration type with the values: + or - ; another *AuthorizationRulePrivileges* element of simpleType of *string* base type with a constraint of enumeration type with the values: *read*, *insert*, *delete*, *update* and *all*; and an *AuthorizationRuleCondition* element of string type that will contain the XPATH expression associated with the expression in OCL.

```
<complexType>
 <sequence>
  <element name= "AuthorizationRuleSign">
   <simpleType>
        <restriction base="string">
          <enumeration value="+/> <enumeration value="-"/>
        </restriction>
   </simpleType>
  </element>
  <element name= "AuthorizationRulePrivileges">
   <simpleType>
        <restriction base="string">
          <enumeration value="read"/>
          <enumeration value="insert"/>
          <enumeration value="delete"/>
          <enumeration value="update"/>
          <enumeration value="all"/>
        </restriction>
   </simpleType>
  </element>
  <element name= "AuthorizationRuleCondition" type="string"/>
 </sequence>
</complexType>
```

According to MDA, once we have applied these rules, the next step is the mapping from PSM to Code of specific DBMSs. These DBMSs usually do not provide security solutions for solving the security issues we consider in our approach, but they support most of XML standards (DOM, XSL, XSLT, XPath, etc.), which allow us to easily implement all these security specifications.

6 Conclusions and Future Work

At the present time, there are different solutions for the storage of XML data but there is no methodology for the systematic design of XML DBs that incorporates security in the development process from its early phases.

In this work, we have integrated the security aspect into the methodological approach for the development of an XML DB in the framework of MIDAS, a model-driven methodology for the development of WIS based on MDA. In the case of the specified development process for secure XML DB, for the secure data PIM, a UML extension to incorporate security aspects at the conceptual level is used. For the secure data PSM we have modified the previously-defined XML DB profile. The incorporation of security aspects has been our main goal. Moreover, we have defined mappings from secure data PIM to secure data PSM that will be the secure XML DB schema. From this logical model of the secure XML DB (PSM), we will obtain the code for the specific XML DB product that we want to use, in a semi-automatic way. Up to now, we have studied the security aspects for the Oracle 10g product, but in future work, we will study other XML DBMSs in detail, in order to analyze which of them take into account security aspects, and how.

A case study for the management of hospital information has been developed, to validate our proposal; we have left this out for the sake of space.

We are now working along several different lines, in an attempt to extend the proposal of this paper. One of these, on which we have already started to work, is the automation of the transformations of the constraints expressed in OCL at the PIM level, to convert them into XPATH language. Moreover, our intention is to automate the transformations between the metamodels and the corresponding models using the incipient Query View Transformation (QVT) proposal [20], which aims to become the standard for defining transformations.

We are also studying the possibility of using XACML [19] as a PSM security rules specification language that could complement the current PSM model (XML Schema). In fact, XACML is a powerful standard language that specifies schemas for authorization policies and for authorization decision requests and response, which is applicable to a wide range of applications, and which can integrate many security policies into a complete security model.

In addition, we want to define queries using the XQuery language, in order to obtain information about the security aspects of the XML DB.

We have a further goal, which is to perform several case studies to detect new needs. These would

also analyze the advantages of incorporating security aspects provided by the different XML DB administrators, not only native ones, but also the XML extensions that DBMSs have. At the same time, we are going to include the security aspect in the subsystem for the semi-automatic development of XML DBs of the tool CASE that we are developing.

Acknowledgements

We would like to thank the reviewers for their valuable comments, which have helped us to improve this paper.

This research has been carried out in the framework of the following projects: GOLD financed by the Spanish Ministry of Education and Science (TIN2005-00010/), DIMENSIONS (PBC-05-012-2) financed by the FEDER and by the "Consejería de Ciencia y Tecnología of the Junta de Comunidades de Castilla-La Mancha" and the RETISTIC network (TIC2002-12487-E) financed by the "Dirección General de Investigación" of the Spanish Ministry of Science and Technology.

References

1. Bertino, E. and Sandhu, R. *Database Security – Concepts, Approaches, and Challenges.* IEEE Transactions on Dependable and Secure Computing. Vol. 2, Nº 1, January-March 2005, pp 2-19, 2005.
2. Bertino, E. and Ferrari, E. *Secure and Selective Dissemination of XML Documents.* ACM Transactions on Information and System Security. Vol. 5, Nº 3, pp. 290-331, 2002.
3. Bertino, E., Castano, S., Ferrari, E., and Mesiti, M. *Specifying and Enforcing Access Control Policies for XML Document Sources*. World Wide Web Journal. Vol. 3. Nº 3, Baltezer Science Publisher, pp. 139-151, 2000.
4. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S. and Samarati, P. *Securing XML Documents.* Proceedings of the 2000 International Conference on Extending Database Technology (EDBT2000), Konstanz, Germany, pp.121-135, 2000.
5. Devanbu, P. and Stubblebine, S. *Software engineering for security: a roadmap*. In: A. Finkelstein (Ed.), The Future of Software Engineering, ACM Press pp. 227-239, 2000.
6. Dhillon, G. and Backhouse, J. *Information System Security Management in the new Millennium.* Communications of the ACM. 43, 7. pp. 125-128, 2000.
7. Fernández-Medina, E. and Piattini M. *Designing secure databases.* Information & Software Technology 47(7), pp. 463-477. 2005
8. Fernández-Medina, E. and Piattini, M. *Extending OCL for Secure Database Design*. In Int. Conference on the Unified Modeling Language (UML 2004). Lisbon (Portugal), October, 2004. Springer-Verlag, LNCS 3273, pp. 380-394. 2004.
9. Fernández-Medina, E., Trujillo, J., Villarroel, R. and Piattini, M. *Extending UML for Designing Secure Data Warehouses.* In Conceptual Modeling (ER 2004). Shanghai (China). November, 2004. Springer Verlag. LNCS 3273, pp. 217-230.
10. Fernández-Medina, E., Trujillo, J., Villarroel, R. and Piattini, M. *Access Control and audit Model for the Multidimensional Modeling of Data Warehouses.* Decision Support Systems. 2006 (In Press).
11. Ferrari E. and Thuraisingham B., *Secure Database Systems*, in: M. Piattini, O. Díaz (Ed.), Advanced Databases: Technology Design. Artech House, 2000.
12. Gabillon, A. and Bruno, E. *Regulating Access to XML Documents.* Proceedings of the 15th Annual IFIP WG 11.3 Working Conference on Database Security, pp. 299-314, 2001.
13. Ghosh, A., Howell C., Whittaker J., *Building software securely from the ground up*, IEEE Software 19 (1) (2002), pp. 14-17, 2002.
14. He, H. and Wong, R.K. *A Role-Based Access Control for XML Repositories.* Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00), 2000.
15. ISACF, Information Security Governance. *Guidance for Boards of Directors and Executive Management, Information Systems Audit and Control Foundation*, USA, 2001.
16. Marcos, E., Vela, B., Cáceres, P. and Cavero, J.M. *MIDAS/DB: a Methodological Framework for Web Database Design.* DASWIS 2001. Yokohama (Japan), November, 2001. Springer-Verlag, LNCS 2465, pp. 227-238, 2002.
17. Marcos, E., Vela, B. and Cavero J.M. *Methodological Approach for Object-Relational Database Design using UML*. Journal on Software and Systems Modeling (SoSyM). Springer-Verlag. Ed.: R. France and B. Rumpe. Vol. SoSyM 2, pp.59-72, 2003.
18. Murata, M., Tozawa, A., Kudo, M. and Hada, S. *XML Access Control Using Static Analysis.* Proceedings of the 10th ACM Conference on Computer and Communication Security, pp.73-84, 2003.
19. OASIS. *eXtensible Access Control Markup Language (XACML 2.0)*. Retrieved from: http://www.oasis-open.org.
20. OMG. *MDA Guide Version 1.0.* Document number omg/2003-05-01. Ed.: Miller, J. and Mukerji, J. Retrieved from: http://www.omg.com/mda, 2003.
21. OMG, Query/Views/Transformation RFP. 2002. Retrieved from: http://omg.org/ad/2002-4-10.
22. Oracle Corporation. *Oracle XML DB. Technical White Paper*. Retrieved from: www.otn.com, 2003.
23. Software AG. *Tamino X-Query. System Documentation Version 3.1.1*. Software AG, Darmstadt, Germany. Retrieved from: www.softwareag.com, 2001.
24. Vela, B., Acuña, C. and Marcos, E. *A Model Driven Approach for XML Database Development*, 23rd. International Conference on Conceptual Modelling (ER2004). Shanghai (China), November, 2004. Springer Verlag, LNCS 3288, pp. 780-794. 2004.
25. Westermann, U. and Klas W. *An Analysis of XML Database Solutions for the Management of MPEG-7 Media Descriptions*. ACM Computing Surveys, Vol. 35 (4), pp. 331-373, 2003.

An Automatic Construction and Organization Strategy for Ensemble Learning on Data Streams

Yi Zhang
School of Software
Tsinghua University, Beijing, 100084 China
zhang-yi@mails.tsinghua.edu.cn

Xiaoming Jin
School of Software
Tsinghua University, Beijing, 100084 China
xmjin@tsinghua.edu.cn

ABSTRACT

As data streams are gaining prominence in a growing number of emerging application domains, classification on data streams is becoming an active research area. Currently, the typical approach to this problem is based on ensemble learning, which learns basic classifiers from training data stream and forms the global predictor by organizing these basic ones. While this approach seems successful to some extent, its performance usually suffers from two contradictory elements existing naturally within many application scenarios: firstly, the need for gathering sufficient training data for basic classifiers and engaging enough basic learners in voting for bias-variance reduction; and secondly, the requirement for significant sensitivity to concept-drifts, which places emphasis on using recent training data and up-to-date individual classifiers. It results in such a dilemma that some algorithms are not sensitive enough to concept-drifts while others, although sensitive enough, suffer from unsatisfactory classification accuracy. In this paper, we propose an ensemble learning algorithm, which: (1) furnishes training data for basic classifiers, starting from the up-to-date data chunk and searching for complement from past chunks while ruling out the data inconsistent with current concept; (2) provides effective voting by adaptively distinguishing sensible classifiers from the else and engaging sensible ones as voters. Experimental results justify the superiority of this strategy in terms of both accuracy and sensitivity, especially in severe circumstances where training data is extremely insufficient or concepts are evolving frequently and significantly.

1. INTRODUCTION

In many emerging applications such as network monitoring, sensor networks, etc., data are produced continually in the form of high-speed streams, which are required to be analyzed on-line. Thus, the applications which aim to classifying data streams rather than static relations are needed. Given the fact that data streams always have the properties such as high-velocity, extremely large volume, and frequently evolving concepts, today's classification techniques meet unprecedented challenges: bounded memory usage, high processing speed, one-pass scanning, any-time available, and so on [4]. Especially, underlying concept of steaming data often alters (termed *concept drift*), which requests that algorithms must be sensitive enough to the up-to-date concept under the data stream [4, 13].

Many strategies have been proposed in order to deal with concept-drifts. For instance, adapting existent models to data streams scenarios [7]; using novel data structure to maintain training data stream and to classify on demand [1]; exhaustively selecting training data by comparing all the sensible choices [5]; or building concept history and combining proactive and reactive modes in prediction [15]. Besides these technologies, the ensemble learning approach [2] appears as a promising solution: it seems reasonable to train individuals to deal with different parts of stream and organize these individual classifiers to make the final decision. This motivates more than a few attempts to develop novel ensemble learning mechanisms for data streams [9, 11, 12, 16]. However, all these models, although effective to some extent, do not provide satisfying solution to some open problems, due to the difficulties of: (1) seeking enough training data for individual classifiers with the guarantee that not importing old concepts; (2) finding adequate voters in global-prediction, while ensuring that experts (i.e. basic classifiers) built upon old concepts are excluded. We discuss these aspects as follows:

Firstly, when building each basic classifier, we want to collect enough data while guarantee that concept-drifts are not imported into training data. To handle this problem, some works split the training data stream into data chunks, and build basic learner from each chunk [11, 12, 16]; while other works use incremental learner as the basic expert, i.e. each expert, after being built, keeps on updating itself until discarded [9]. In fact, both of these two methods can not furnish ideal solution. On the one hand, fixing the amount of training data for basic classifier by size of chunk is questionable. Given the fact that the velocity of training data stream is often limited by the manual labeling process, the size of data chunk can not be very large because large chunk needs relatively long period to be accumulated, thus leads to high possibility that concept-drift happens in this period. Nonetheless, if basic classifiers can not obtain sufficient training data, the ensemble will not work effectively. On the other hand, using incremental classifier

also suffers from some flaws. It is true that allowing each individual expert to adjust itself according to future training data is beneficial to this individual [9]. But this approach has negative effects on the whole ensemble: when an old learner is incompatible with the latest concept, the most optimal policy is discarding it and allowing the "right ones" to make decision, rather than adjusting (if possible) the elder, which actually postpones its retirement. Moreover, though incremental learning gives the individual the chance for improving itself, the bias can not be completely corrected in that the learner is built from old data and merely "update" itself based on newcome data.

Secondly, when using basic classifiers to form global predictor, we want to engage adequate voters in final decision for the sake of bias-variance reduction [2], while ensure that outmoded classifiers are obviated. Although recent works place much stress on this point, none of them can make good balance. In [11], the global prediction is made by majority voting among N "high quality" individuals. The drawback of this method is clear-cut: Only after more than $N/2$ members in ensemble mastery the new concept (which needs at least $N/2$ new data chunks after concept-drift occurs), the majority voting will make correct prediction. Thereafter, some works focus on improving voting's sensitivity to concept-drift [9, 12]. For example in [9]: (1) the ensemble is composed of classifiers whose "quality" larger than an threshold q_0 rather than uses fixed amount of basic classifiers; (2) The global prediction is based on weighted voting rather than majority voting. Although this approach improves ensemble's sensitivity to concept-drift, it still has problems. First of all, q_0 is difficult to choose: we want good voters, but we also need enough voters. Second of all, weight-based voting can not eliminate the negative effect of out-of-date experts ---- they still can overwhelm the sensible ones by larger total weight. Since neither of majority voting and weight-based voting can produce sensitive ensemble, the "apparently" substituted way is "trusting in" the best rather than voting by the masses [16]. Whereas, simply engaging the best classifier will lose important advantage of voting-based ensemble: bias and variance reduction [2]. In fact, when using some unstable learners such as C4.5 [10], voting-based ensemble such as bagging can improve the accuracy by dramatically reducing variance [2, 3]. Even for stable classifiers such as naïve Bayes [8], voting strategy as boosting [6] has positive effect by decreasing bias [2].

In this paper, we propose a dynamic ensemble learning algorithm, termed Dynamic Construction and Organization (DCO), which concentrates on these two difficulties. The contributions and key ideas of this work are: (1) the *individual-construction strategy* which provides training data for basic classifiers, starting from the latest data chunk and searching complement from history while excluding the data inconsistent with current concept; (2) the *global-prediction policy* which offers effective voting by adaptively differentiating between sensible experts and the else and engaging sensible ones as voters. Experimental results show that our ensemble approach achieves high accuracy and remains sensitivity to concept-drifts.

This paper is organized as follows. Section 2 describes our approach, section 3 provides the experimental results, and section 4 concludes the paper.

2. Dynamic Construction and Organization Strategy for Ensemble Learning

In this section, we put forward our DCO (Dynamic Construction and Organization) approach. After introducing the problem definition and framework of the algorithm, we mainly focus on the individual-construction and global-prediction strategies. It is assumed that training data and testing data are given as data streams, termed S and T in our paper, respectively. Data items in S are divided into data chunks, with size of *chunkSize*. As a rule, we set the latest chunk from S as evaluation dataset. When future chunk is available, current evaluation dataset can be used as training chunk and the coming chunk is set as new evaluation dataset. The algorithm framework is: (1) when a new training chunk is available, we use *individual-construction* strategy to create a new basic classifier from this chunk plus the old chunks; (2) we set the most recent N basic classifiers as the ensemble; (3) for each test point, we use the ensemble to classify the data based on *global-prediction* strategy.

2.1 Individual-Construction Strategy

Table 2 shows our ***Individual-Construction Strategy*** which pursues a balance between data sufficiency and sensitivity, especially when single chunk is not enough for training basic learner. Function ***create*** is depend on the basic learner. In this paper, we have tested both C4.5 [10] and naïve Bayes [8], see section 3 for details. What is more, there are two additional functions, ***dataSelect*** and ***outperform***, discussed in following subsections.

Table 2. Individual-construction strategy

Input:
  $D_n, D_{n-1}, \ldots D_1$: data chunks available
Output:
  C_n: resulting new expert
Variable:
  D: training data for new basic learner
  $\triangle$:selected data from old chunk
  C_n': alternative expert

$D \leftarrow D_n$
$C_n \leftarrow$ ***create*** (D)
for $i = n-1$ to 1
  $\triangle \leftarrow$ ***dataSelect*** (D_i)
  $C_n' \leftarrow$ ***create*** $(D+\triangle)$

```
        if outperform(C_n', C_n )
            C_n ← C_n'
            D ← D+△
        else
            return C_n
        end-if
    end-for
    return C_n
```

2.1.1 Data Selection Function

This function aims at selecting complementary data for D. Here we assume no concept-drift in D_i (*outperform* will deal with concept-drift). But even under stationary concept, unselectively importing old data is harmful because (1) it makes the learner over-fit the old part; (2) unnecessarily large amount of training data slows down the learning. In this sense, we define the *dataSelect* as choosing: (1) data in D_i that are misclassified by C_n, plus (2) data that are misclassified by previous learner C_{n-1}. Choosing data misclassified by C_n is based on the hypothesis that C_n has not mastered this part of data and thus needs further learning. The idea of importing data misclassified by C_{n-1} is inspired by Boosting [2, 6]: each learner puts emphasis on the "difficult" part for its predecessor. From this perspective, *dataSelect* may bring additive benefits in two aspects [2, 6]: (1) reducing bias; (2) augmenting the diversity among individuals. Both of these will improve the performance of ensemble.

2.1.2 Evaluation Function

Outperform evaluates C_n and C_n', and makes decision that whether importing $\triangle$ to D is sensible. Since $\triangle$ is made up of misclassified data, we must be wary of two possibilities: (1) Misclassification caused by concept-drift; (2) Misclassification caused by noise. In these two cases, introducing such misclassified data will do harm to training. Furthermore, when improvement is insignificant, importing should also be stopped for the sake of efficiency.

The process for evaluating C_n and C_n' is as follows: Firstly, compute the prediction accuracy of C_n and C_n' (termed p and p', respectively) based upon evaluation dataset. Secondly, calculate lower-bound (termed *low* and *low'*) for p and p' under confidence *conf*, according to equation (1). Thirdly, if and only if $low' - low > \varepsilon$ holds for threshold ε, we judge that C_n' *outperform* C_n.

$$low = \left(p + \frac{z^2}{2V} - z\sqrt{\frac{p}{V} - \frac{p^2}{V} + \frac{z^2}{4V^2}} \right) \Big/ \left(1 + \frac{z^2}{V} \right) \quad (1)$$

In equation (1), z satisfies $P(X \geq z) = conf$ under normal distribution and $V = chunkSize$. The intuition of this equation is: given a prediction accuracy p based on a test set of size V, we assume p is a random variable that has mean m and standard deviation $\sqrt{m(1-m)/V}$. Then (2) holds, which naturally leads to (1) where *low* is one solution of m.

$$P\left(-z < (p-m) \Big/ \sqrt{m(1-m)/V} < z \right) = conf \quad (2)$$

2.2 Global-Prediction Strategy

In section 1, we have reviewed different policies to organize global predictor, such as majority voting, weight-based voting and select-best. In fact, the ideal strategy should strike a balance between these choices. On one hand, it should retain the benefits of voting by masses rather than simply select the best individual. On the other hand, we want the sensible experts to dominate the voting, thus render the global predictor sensitive to concept-drift.

2.2.1 Dynamic Voting

Our strategy is based upon the fact that we only want to divide the ensemble into two categories: the "good enough" experts and the else. Since we assume merely two categories in basic learners, it is reasonable to expect certain simple method to "judge good and evil in such a melodrama". Here we put forward an efficient procedure to choose voters from ensemble.

(1) Sort N basic classifiers in ensemble according to their accuracies on evaluation dataset.
(2) Among N-1 distances between sorted classifiers, find the maximal one.
(3) The maximal distance naturally divides the learners into two groups.
(4) Engage the "better" group as voting group.

The time complexity of this procedure depends on sorting step, which is trivial when N only refers to the capacity of ensemble. Furthermore, this procedure is executed only when the evaluation dataset is replaced by new chunk (the ensemble will be updated at the same time). Based on this voting policy, choosing ensemble capacity N is easy ---- we can choose a larger quantity than other voting-based algorithms, for the reason that outmoded experts in ensemble will be excluded from voting group by our dynamic voting. It will benefit in two aspects: (1) Under stationary concept, large ensemble furnishes sufficient voters; (2) In concept-drift scenario, large ensemble offers more opportunities for finding sensible experts, especially when concept switches in a repeated way.

2.2.2 Discussion: Other Choices?

Now we discuss that whether some other simper strategies can be used instead of our dynamic voting: (1) "select best-k": For N experts in ensemble, select k best experts as voters. (2) "Performance threshold": according to a threshold p_0, define experts in ensemble whose

accuracies higher than p_0 as voters. Firstly, the "select best-k" policy aims at retaining the sensitivity of "select-best" policy and gaining the benefits of voting. Nonetheless, this strategy is obviously incompetent in that it is actually the similar with "majority voting" where $N = k$, whose flaws have been discussed in Section 1. Secondly, the "performance threshold" is not an ideal approach, either. In fact, we can not decide this threshold in order to divide the ensemble into "sensible" ones and the else: (1) Performance of basic classifier changes dramatically on different classification problems. (2) It is unknown that to what extent the concept-drift will degrade the performance of outmoded experts and where should we set this threshold.

3. Empirical Study and Results

This section presents the results of our experimental evaluation of the proposed method. The goal of our experiments is to demonstrate the ability of our algorithm to: (1) handle data insufficiency when training basic classifiers; (2) form effective voting; (3) keep sensitive to concept-drifts.

3.1 Dataset and System Implementation

To determine the performance of our algorithm on problems involving concept-drifts, we design the problem in which each data points has three attributes $x, y, z \in R$, randomly sampled from range [0, 10]. The data point that satisfies the target concept $x^2 + y^2 + z^2 < r^2$ is labeled by 1. Otherwise the item will be labeled as 0. Radius r is used to control the concept-drifts. Experiments are implemented on Weka toolkit [14].

3.2 Concept-drift Tests and Results

Four algorithms are tested: (1) SEA: algorithm in [11]; (2) DWM: algorithm in [9]; (3) DCS: algorithm in [16]; (4) DCO: our algorithm. DWM does not take part in *Test1* and *Test2* since it must use incremental basic classifier. All results are averaged from 30 independent runs.

(1) *Test1*: Testing SEA, DCS, and DCO based on C4.5, *chunkSize* = 50.

(2) *Test2*: Testing SEA, DCS, and DCO based on C4.5, *chunkSize* = 100.

(3) *Test3*: Testing SEA, DCS, DCO and DWM based on Naïve Bayes, *chunkSize* = 50.

(4) *Test4*: Testing SEA, DCS, DCO and DWM based on Naïve Bayes, *chunkSize* = 100.

The procedure of experiment is: There are entirely 50 *chunkSize* training data points. For the first fourth the radius r in target concept is 9; for the second r = 11.5; for the third r = 8.5; for the last r = 11. For each fourth, we randomly generate a testing dataset of 2500 data points on corresponding radius. Each time after *chunkSize* training

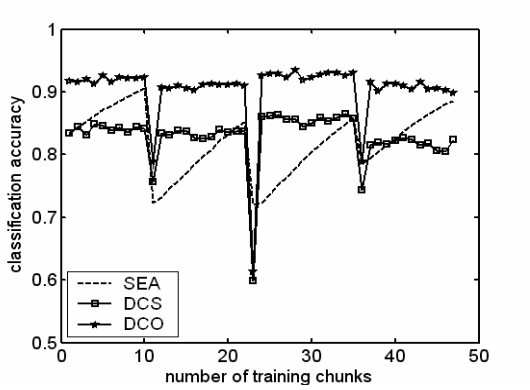

Figure 1: Results of *test 1*.

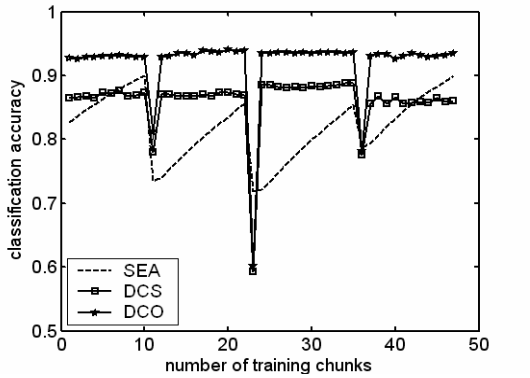

Figure 2: Results of *test 2*.

data points are offered, we test all the algorithms using appropriate testing dataset. For our algorithm, *conf* = 0.9 and ε = 1% in *outperform* function. For all algorithms with fixed-size ensemble, we set N = 50. Other parameters are set according to original papers. See Fig.1~Fig.4 for results, the analysis of these results is as follows:

(1) **Prediction accuracy:** DCO has the best classification accuracy, and this advantage appears more evident when the size of data chunk is limited (*chunkSize* = 50). Such superiority dues great part to our novel strategies for individual-construction and global-prediction. In one sense, the former policy guarantees the sufficiency of training data for basic learners, augments the diversity among individuals, and reduces the bias of basic learners. In another sense, the latter strategy strikes a balance between the quantity of voters and the quality of voters, and thus renders the voting process much more effective in terms of variance-reduction.

(2) **Sensitivity for concept-drift:** DCO and DCS are quite sensitive to concept-drift: they recover from misclassification very fast; DWM is not as sensitive as DCO and DCS, but still much better than SEA. DCS's sensitivity obviously dues to its "select-best" policy; DCO relies on dynamic voting to exclude outmoded experts, thus remains as sensitive as DCS; DWM uses weighted-based

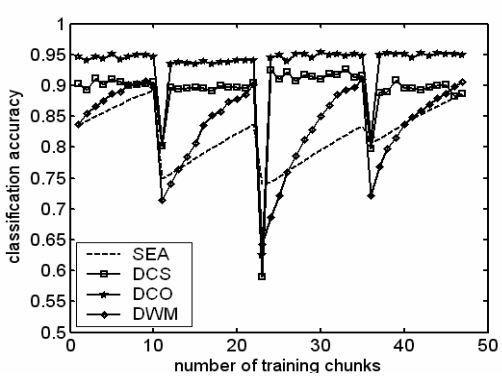

Figure 3: Results of *test 3*.

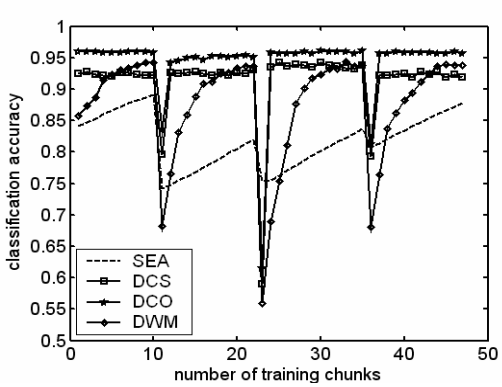

Figure 4: Results of *test 4*.

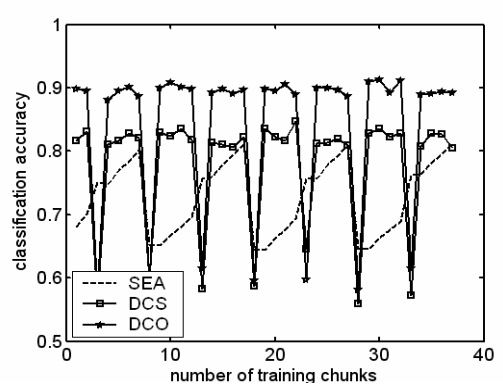

Figure 5: Results of *test 5*.

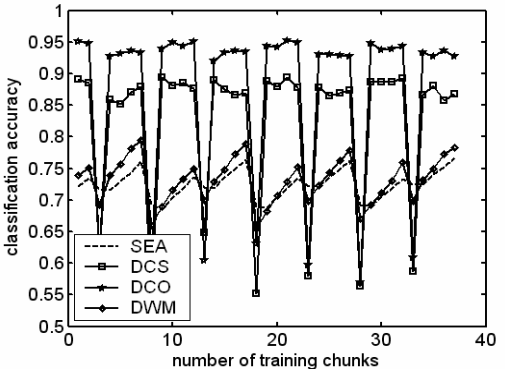

Figure 6: Results of *test 6*.

voting, and does not fix the capacity of ensemble, therefore has better alertness than SEA, which engages majority voting on fixed amount of voters in ensemble.

(3) **DCS and basic learner:** DCS algorithm performed much better under naïve Bayes than using C4.5, because the former is a stable basic learner which is not in dire need of voting to reduce its variance. However, using unstable classifiers such as C4.5, DCS will appear ineffective. Furthermore, DCO beats DCS even based on naïve Bayes, since DCO enhances data sufficiency, individual diversity, and further reduces the bias-variance by dynamic voting.

(4) **Efficiency of algorithms:** We test the efficiency of the four algorithms, represented by the time consumed in their 30 independent runs and shown in Table 1. SEA is time-consuming, especially when using Naïve Bayes. DCO is as efficient as DCS based on C4.5, and retains a reasonable speed on Naïve Bayes. In fact, the most complex part of DCO, the individual construction process mentioned in section 2.1, always stops after combining a few old blocks. Note that DWM used more time in test3 than in test4 since small data blocks lead to frequent creation of new classifers.

3.3 Performance in Severe Circumstance

What is more interesting is the performance of these algorithms in severe conditions: data insufficiency plus frequent and sudden concept-drifts, where data insufficiency calls for the ability to seek enough data for training basic classifiers; and frequent abrupt concept-drifts require excluding old concepts from training data. We set *chunkSize* as 25. For totally 40 *chunkSize* training data points, concept-drift happens after each 4 chunks. The radius starts with r=8, switches between 8 and 12 (i.e. r = 8, 12, 8, 12 …). For each concept, we randomly generate 2500 data points by corresponding radius. Each time after a chunk of training data is offered, we test all the algorithms using appropriate testing points. Other settings are similar with section 3.2. *Test5* concerns SEA, DCS, and DCO based on C4.5; *test6* measures SEA, DCS, DCO and DWM on naïve Bayes. The results are shown in Fig. 5 and Fig. 6. We can observe that: (1) the individual-construction policy effectively handles the data insufficiency; (2) the dynamic voting strategy furnishes successful voting, while retains the sensitivity to sudden and frequent concept-drifts.

Table 1. Efficiency of Algorithms

	SEA	DCS	DCO	DWM
Test1	6min 11sec	1min 4sec	1min 39sec	---
Test2	8min 1sec	1min 46sec	1min 51sec	---
Test3	51min 9sec	3min 7sec	14min 9sec	11min 5sec
Test4	78min 4sec	5min 21sec	15min 3sec	6min 1sec

3.4 Real-world Dataset

In this section, we proposed our empirical results on "adult" dataset [17]. We test SEA, DCS and DCO upon C4.5. The training and testing dataset contain 32561 and 16281 instances, respectively. Data has 14 attributes such as the age, occupation and sex of a person. The label indicates whether this person has an income larger than 50k dollar. The preprocessing step aims to produce sufficient concept drifts: partition the dataset by "occupation" attribute, then collect instances in three occupations – "Adm-clerical", "Exec-managerial" and "Other-service"; finally we get three training subsets with 3770, 4066 and 3295 instances and three test subsets with 1841, 2020 and 1628 instances, respectively. We set blockSize as 100. Totally 90 data blocks are engaged for training: 15 blocks from subset1, 15 blocks from subset2, 15 blocks from subset3, and then repeat. After each block, we test all the algorithms using corresponding test dataset. For all algorithms, ensemble size is 30. The results are shown in Table 2, which justifies the superiority of DCO.

Table 2. Empirical Results on "Adult" Dataset

SEA	DCS	DCO
0.7773	0.8401	0.8510

4. Conclusions

Current algorithms for mining data streams are confronted with two contradictory elements: Firstly is the need for seeking adequate training data for each basic classifier and gathering sufficient voters for final-decision; and secondly, is the requirement for sensitivity to concept-drift, which calls for using recent training data and up-to-date basic classifier. In this work, we initially point out the essential reasons for the incompetence of several recent algorithms in solving these conflicting elements. Then, we propose a dynamic ensemble learning algorithm, termed DCO (Dynamic Construction and Organization), which aims at reconciling these contradictions. Experimental results justify the superiority of our approach over the state-of-the-art algorithms in that individual-construction strategy provides solution to data insufficiency under concept-drift scenario; and the dynamic voting strategy strikes a balance between the quantity and quality of voters.

5. Acknowledgement

This work is supported by the National Science Foundation of China (60403021) and the 973 Program (2004CB719400).

6. References

[1] CC Aggarwal, J Han, J Wang, PS Yu. On Demand Classification of Data Streams. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[2] E. Bauer, R. Kohavi. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning, vol 36, pp 105-139, 1999.

[3] L. Breiman. Bagging Predictors. Machine Learning, vol 24, pp 123-140, 1996.

[4] GZ Dong, JW Han, Laks V.s. Lakshmanan, J Pei, HX Wang, Philip S. Yu. Online Mining of Changes from Data Streams: Research Problems and Preliminary Results. ACM SIGMOD MPDS`03 San Diego, CA, USA.

[5] W Fan. Systematic Data Selection to Mine Concept-Drifting Data Streams. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[6] Y Freund, RE Schapire. Experiments with a New Boosting Algorithm. Machine Learning: Proceedings of the 13th International Conference, 1996.

[7] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001.

[8] G. H. John and P. Langley. Estimating Continuous Distributions in Bayesian Classifiers. In Proc. of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, 1995, 338-345.

[9] JZ Kolter, MA Maloof. Dynamic Weighted Majority: A New Ensemble Method for Tracking Concept Drift. Proceedings of the Third IEEE International Conference on Data Mining, 2003.

[10] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.

[11] W. N. Street, YS Kim. A streaming ensemble algorithm (SEA) for large-scale classification. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, ACM Press, 2001, 377-382.

[12] H Wang, W Fan, PS Yu, J Han. Mining concept-drifting data streams using ensemble classifiers. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

[13] G. Widmer and M. Kubat. Learning in the Presence of Concept Drift and Hidden Contexts. Machine Learning, vol23, issue1, 1996, 69-101.

[14] I. H. Witten and E. Frank. 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Mateo, CA.

[15] Y Yang, X Wu, X Zhu. Combining Proactive and Reactive Predictions for Data Streams. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005.

[16] XQ Zhu, XD Wu and Y Yang. Dynamic Classifier Selection for Effective Mining from Noisy Data Streams. In: Proc. 4th IEEE Int'l Conf. on Data Mining, 2004, 305-312.

[17] http://www.ics.uci.edu/~mlearn/MLRepository.html

A Survey on Ontology Mapping

Namyoun Choi, Il-Yeol Song, and Hyoil Han
College of Information Science and Technology
Drexel University, Philadelphia, PA 19014

Abstract

Ontology is increasingly seen as a key factor for enabling interoperability across heterogeneous systems and semantic web applications. Ontology mapping is required for combining distributed and heterogeneous ontologies. Developing such ontology mapping has been a core issue of recent ontology research. This paper presents ontology mapping categories, describes the characteristics of each category, compares these characteristics, and surveys tools, systems, and related work based on each category of ontology mapping. We believe this paper provides readers with a comprehensive understanding of ontology mapping and points to various research topics about the specific roles of ontology mapping.

Introduction

"An ontology is defined as a formal, explicit specification of a shared conceptualization." [27] Tasks on distributed and heterogeneous systems demand support from more than one ontology. Multiple ontologies need to be accessed from different systems. The distributed nature of ontology development has led to dissimilar ontologies for the same or overlapping domains. Thus, various parties with different ontologies do not fully understand each other. To solve these problems, it is necessary to use ontology mapping geared for interoperability. This article aims to present the broad scope of ontology mapping, mapping categories, their characteristics, and a comprehensive overview of ontology mapping tools, systems, and related work.

We classify ontology mapping into the following three categories: 1) mapping between an integrated global ontology and local ontologies [3, 4, 1, 7], 2) mapping between local ontologies [6, 1, 8, 9, 12, 13, 14], and 3) mapping on ontology merging and alignment.[15, 16, 17, 18, 19, 20]

The first category of ontology mapping supports ontology integration by describing the relationship between an integrated global ontology and local ontologies. The second category enables interoperability for highly dynamic and distributed environments as mediation between distributed data in such environments. The third category is used as a part of ontology merging or alignment as an ontology reuse process.

In this paper, we survey the tools, systems, and related work about ontology mapping based on these three ontology mapping categories. A comparison of tools or systems about ontology mapping is made based on specific evaluation criteria[10], which are input requirements, level of user interaction, type of output, content of output, and the following five dimensions: structural, lexical, domain, instance-based knowledge, and type of result.[8] Through a comparative analysis of ontology mapping categories, we aim to provide readers with a comprehensive understanding of ontology mapping and point to various research topics about the specific roles of ontology mapping.

The paper is organized as follows. The meanings of ontology mapping[4, 3, 7, 15, 25], ontology integration, merging, and alignment [2, 24] are outlined in Section 2. In Section 3, characteristics and application domains of three different categories of ontology mapping are discussed. The tools, systems, frameworks, and related work of ontology mapping are surveyed based on the three different ontology mapping categories. Then the overall comparison of tools or systems about ontology mapping is presented. In Section 4, a conclusion and presentation of future work are detailed.

2. Terminology: ontology mapping, ontology integration, merging, and alignment

In this section, we set the scope of ontology mapping and ontology mapping tools, and outline meanings of ontology mapping, integration, merging, and alignment. We aim to give a wide view of ontology mapping including ontology integration, merging, and alignment because this concept of ontology mapping is broad in scope[5] and ontology mapping is required in the process of ontology integration, merging, and alignment. Furthermore, one closely related research topic with ontology mapping is schema matching, which has been one major area of database research.[3, 36, 37, 38] However, this is beyond our scope in this paper. We also refer to tools for ontology integration, merging, and alignment as ontology mapping tools in this paper. We discuss the meanings of ontology mapping based on the three different ontology mapping categories.

Ontology merging, integration, and alignment

Ontology merging, integration, and alignment can be considered as an ontology reuse process.[2,24]

Ontology merging is the process of generating a single, coherent ontology from two or more existing and different ontologies related to the same subject.[26] A merged single coherent ontology includes information from all source ontologies but is more or less unchanged. The original ontologies have similar or overlapping domains but they are unique and not revisions of the same ontology.[24]

Ontology alignment is the task of creating links between two original ontologies. Ontology alignment is made if the sources become consistent with each other but are kept separate.[15] Ontology alignment is made when they usually have complementary domains.

Ontology integration is the process of generating a single ontology in one subject from two or more existing and different ontologies in different subjects.[26] The different subjects of the different ontologies may be related. Some change is expected in a single integrated ontology. [26]

Ontology mapping

Ontology mapping between an integrated global ontology and local ontologies.[4, 3, 7] In this case, ontology mapping is used to map a concept found in one ontology into a view, or a query over other ontologies (e.g. over the global ontology in the local-centric approach, or over the local ontologies in the global-centric approach).

Ontology mapping between local ontologies.[25] In this case, ontology mapping is the process that transforms the source ontology entities into the target ontology entities based on semantic relation. The source and target are semantically related at a conceptual level.

Ontology mapping in ontology merge and alignment.[15] In this case, ontology mapping establishes correspondence among source (local) ontologies to be merged or aligned, and determines the set of overlapping concepts, synonyms, or unique concepts to those sources.[15] This mapping identifies similarities and conflicts between the various source (local) ontologies to be merged or aligned.[5]

3. Categories of Ontology Mapping

In this section, ontology mapping based on the following three categories will be examined: 1) ontology mapping between an integrated global ontology and local ontologies, 2) ontology mapping between local ontologies, and 3) ontology mapping in ontology merging and alignment.

One of the crucial differences among the three ontology mapping categories is how mapping among ontologies is constructed and maintained. Each category of ontology mapping has different characteristics (strengths and drawbacks). Ontology mapping plays an important role in different application domains[5] and is the foundation of several applications.[14]

3.1 Ontology mapping between an integrated global ontology and local ontologies

This category supports ontology integration processes. Methodological aspects of ontology integration relate to how this mapping is defined.[1] This mapping specifies how concepts in global and local ontologies map to each other, how they can be expressed based on queries[7], and how they are typically modeled as views or queries (over the mediated schema in the local-as-view approach, or over the source schemas in the global-as-view approach).[7]

3.1.1 Strengths and drawbacks

The strengths of this mapping can also be the drawbacks of mapping between local ontologies and vice versa. In this mapping, it is easier to define mapping and find mapping rules than in mapping between local ontologies because an integrated global ontology provides a shared vocabulary and all local ontologies are related to a global ontology. It can be difficult to compare different local ontologies because no direct mappings exist between local ontologies. This mapping lacks maintainability and scalability because the change of local ontologies or the addition and removal of local ontologies could easily affect other mappings to a global ontology. This mapping requires an integrated global ontology. But there exists a practical impossibility of maintaining it in a highly dynamic environment.[8] This mapping cannot be made among different ontologies which have mutually inconsistent information over the same domain or over a similar view of domain because a global ontology cannot be created.

3.1.2 Application domains

This mapping supports the integration of ontologies for the Semantic Web, enterprise knowledge management, and data or information integration. In the Semantic Web, an integrated global ontology extracts information from the local ones and provides a unified view through which users can query different local ontologies.[7] When managing multiple ontologies for enterprise knowledge management, different local ontologies (data sources) can be combined into an integrated global ontology for a query.[1] In an information integration system, a mediated schema is constructed

for user queries. Mappings are used to describe the relationship between the mediated schema (i.e., an integrated global ontology) and local schemas.[1,7,3,4] Ontology is more complicated and expressive in semantics than schema and has some differences but shares many features.[34, 35, 5] Schema can still be viewed as an ontology with restricted relationship types.[9] Therefore, the mediated schema can be considered as a global ontology.[3]

3.1.3 Tools, systems, and related work

An integrated global ontology (the logical mediated schema) is created as a view.[4,7,3] Mappings are used to describe the relationship between the mediated schema and local schemas.

LSD[3] (Learning Source Description): LSD semi-automatically creates semantic mappings with a multi-strategy learning approach. This approach employs multiple learner modules with base learners and the meta-learner where each module exploits a different type of information in the source schemas or data. LSD uses the following base learners: 1) The Name Learner: it matches an XML element using its tag name, 2) The Content Learner: it matches an XML element using its data value and works well on textual elements, 3) Naïve Bayes Learner: it examines the data value of the instance, and doesn't work for short or numeric fields, and 4) The XML Learner: it handles the hierarchical structure of input instances. Multi-strategy learning has two phases: training and matching. In the training phase, a small set of data sources has been manually mapped to the mediated schema and is utilized to train the base learners and the meta learner. In the matching phase, the trained learners predict mappings for new sources and match the schema of the new input source to the mediated schema. LSD also examines domain integrity constraints, user feedback, and nested structures in XML data for improving matching accuracy. LSD proposes semantic mappings with a high degree of accuracy by using the multi-strategy learning approach.

MOMIS[4] (Mediator Environment for Multiple Information Sources): MOMIS creates a global virtual view (GVV) of information sources, independent of their location or their data's heterogeneity. MOMIS builds an ontology through five phases as follows:

1) Local source schema extraction by wrappers
2) Local source annotation with the WordNet
3) Common thesaurus generation: relationships of inter-schema and intra-schema knowledge about classes and attributes of the source schemas
4) GVV generation: A global schema and mappings between the global attributes of the global schema and source schema by using the common thesaurus and the local schemas are generated.
5) GVV annotation is generated by exploiting annotated local schemas and mappings between local schemas and a global schema.

MOMIS generates mappings between global attributes of the global schema and source schemas. For each global class in the global virtual view (GVV), a mapping table (MT) stores all generated mappings. MOMIS builds an ontology that more precisely represents domains and provides an easily understandable meaning to content, a way to extend previously created conceptualization by inserting a new source.

A Framework for OIS[7] (Ontology Integration System): Mappings between an integrated global ontology and local ontologies are expressed as queries and ontology as Description Logic. Two approaches for mappings are proposed as follows: 1) concepts of the global ontology are mapped into queries over the local ontologies (global-centric approach), and 2) concepts of the local ontologies are mapped to queries over the global ontology (local-centric approach).

3.2 Ontology mapping between local ontologies

This category provides interoperability for highly dynamic, open, and distributed environments and can be used for mediation between distributed data in such environments.[12] This mapping is more appropriate and flexible for scaling up to the Web than mappings between an integrated global ontology and local ontologies.[12]

3.2.1 Strengths and drawbacks

This mapping enables ontologies to be contextualized because it keeps its content local.[6] It can provide interoperability between local ontologies when different local ontologies cannot be integrated or merged because of mutual inconsistency of their information.[6,1] It is useful for highly dynamic, open, and distributed environments[6] and also avoids the complexity and overheads of integrating multiple sources.[1] Compared to mapping between an integrated ontology and local ontologies, this category mapping has more maintainability and scalability because the changes (adding, updating, or removing) of local ontology could be done locally without regard to other mappings. Finding mappings between local ontologies may not be easier than between an integrated ontology and local ontologies because of the lack of common vocabularies.

3.2.2 Application domains

The primary application domains of this mapping are the Web or the Semantic Web because

of their de-centralized nature. When there is no central mediated global ontology and coordination has to be made using ontologies, then mappings between local ontologies are necessary for agents to interoperate.[14] In distributed knowledge management systems, when building an integrated view is not required or multiple ontologies cannot be integrated or merged because of mutual inconsistency of the information sources, this category of mapping is required between local ontologies.[1,6]

3.2.3 Tools, systems, and related work

Context OWL[6] (Contextualizing Ontologies): OWL syntax and semantics are extended. Ontologies cannot be integrated or merged as a single ontology if two ontologies contain mutually inconsistent concepts. However, those two ontologies can be mapped using bridge rules which are the basic notion about the definition of context mappings.[6] A mapping between two ontologies is a set of bridge rules using $\supseteq$, $\subseteq$, $\equiv$, $*$ (related), and $\perp$ (unrelated).

CTXMATCH[8]: CTXMATCH is an algorithm for discovering semantic mappings across hierarchical classifications (HCs) using logical deduction. CTXMATCH takes two inputs H, and H1 in HCs, and for each pair of concepts $k \in H$, $k1 \in H1$ (a node with relevant knowledge including meaning in Hierarchical classifications), returns their semantic relation ($\supseteq$, $\subseteq$, $\equiv$, $*$, and $\perp$). For example, k is more general than k1 ($k \supseteq k1$), k is less general than k1 ($k \subseteq k1$), k is equivalent to k1 ($k \equiv k1$), k is compatible with k1 ($k * k1$), and k is incompatible with k1 ($k \perp k1$).

The contribution of the CTXMTCH is that mappings can be assigned a clearly defined model-theoretic semantics and that structural, lexical, and domain knowledge are considered.

GLUE[9]: GLUE semi-automatically creates ontology mapping using machine learning techniques. GLUE consists of Distribution Estimator, Similarity Estimator, and Relaxation Labeler. GLUE finds the most similar concepts between two ontologies and calculates the joint probability distribution of the concept using a multi-strategy learning approach for similarity measurement. GLUE gives a choice to users for several practical similarity measures. GLUE has a total of three learners: Content Learner, Name Learner, and Meta Learner. Content and Name Learners are two base learners, while Meta Learner combines the two base learners' prediction. The Content Learner exploits the frequencies of words in content of an instance (concatenation of attributes of an instance) and uses the Naïve Bayes' theorem. The Name Learner uses the full name of the input instance. The Meta-Learner combines the predictions of base learners and assigns weights to base learners based on how much it trusts that learner's predictions. In GLUE, Relaxation Labeling takes a similarity matrix and reaches for the mapping (best label assignment between nodes (concepts)). This mapping configuration is the output of GLUE.

MAFRA[12] (Ontology **MA**pping **FRA**mework for distributed ontologies in the Semantic Web): MAFRA provides a distributed mapping process that consists of five horizontal and four vertical modules.[12] Five horizontal modules are as follows:

1) Lift & Normalization: It deals with language and lexical heterogeneity between source and target ontology.
2) Similarity Discovery: It finds out and establishes similarities between source ontology entities and target ontology entities.
3) Semantic Bridging: It defines mapping for transforming source instances into the most similar target instances.
4) Execution: It transforms instances from the source ontology into target ontology according to the semantic bridges.
5) Post-processing: It takes the result of the execution module to check and improve the quality of the transformation results.

Four vertical modules are as follows:

1) Evolution: It maintains semantic bridges in synchrony with the changes in the source and target ontologies.
2) Cooperative Consensus Building: It is responsible for establishing a consensus on semantic bridges between two parties in the mapping process.
3) Domain Constraints and Background Knowledge: It improves similarity measure and semantic bridge by using WordNet or domain-specific thesauri.
4) Graphical User Interface (GUI): Human intervention for better mapping.

MAFRA maps between entities in two different ontologies using a semantic bridge, which consists of concept and property bridges. The concept bridge translates source instances into target ones. The property bridge transforms source instance properties into target instance properties.

LOM[21] (Lexicon-based Ontology Mapping): LOM finds the morphism between vocabularies in order to reduce human labor in ontology mapping using four methods: whole term, word constituent, synset, and type matching. LOM does not guarantee accuracy or correctness in mappings and has limitations in dealing with abstract symbols or codes in chemistry, mathematics, or medicine.

QOM[22] (Quick Ontology Mapping): QOM is a efficient method for identifying mappings between two ontologies because it has lower run-time complexity. In order to lower run-time complexity

QOM uses a dynamic programming approach.[33] A dynamic programming approach has data structures which investigate the candidate mappings, classify the candidate mappings into promising and less promising pairs, and discard some of them entirely to gain efficiency. It allows for the ad-hoc mapping of large-size, light-weight ontologies.

ONION[13] (**ON**tology composit**ION** system): ONION resolves terminological heterogeneity in ontologies and produces articulation rules for mappings. The linguistic matcher identifies all possible pairs of terms in ontologies and assigns a similarity score to each pair. If the similarity score is above the threshold, then the match is accepted and an articulation rule is generated. After the matches generated by a linguistic matcher are available, a structure-based matcher looks for further matches. An inference-based matcher generates matches based on rules available with ontologies or any seed rules provided by experts. Multiple iterations are required for generating semantic matches between ontologies. A human expert chooses, deletes, or modifies suggested matches using a GUI tool. A linguistic matcher fails when semantics should be considered.

OKMS[1] (Ontology-based knowledge management system): OKMS is an ontology-based knowledge management system. In OKMS, mapping is used for combining distributed and heterogeneous ontologies. When two different departments deal with the same business objects, their ontologies for their systems do not match because they approach the domain from different perspective. When they want to include information from other departments in their knowledge management system, the information must be transformed (i.e., reclassified). This can be accomplished through a mapping between local ontologies. The five-step ontology-mapping process[12] is used in the OKMS. The five-step ontology mapping process is as follows: 1) Lift and normalization: If source information is not ontology-based, it will be transformed to the ontology level by a wrapper. 2) Similarity extraction: The similarity extraction phase creates a similarity matrix, which represents the similarities between concepts and instances in ontologies being mapped. 3) Semantic mapping: This step produces the mappings that define how to transform source-ontology instances into target-ontology instances. 4) Execution: Execute mappings. 5) Post-processing: It improves the results of the execution phase.

OMEN[31] (Ontology Mapping Enhancer): OMEN is a probabilistic ontology mapping tool which enhances the quality of existing ontology mappings using a Bayesian Net. The Bayesian Net uses a set of meta-rules that represent how much each ontology mapping affects other related mappings based on ontology structure and the semantics of ontology relations. Existing mappings between two concepts can be used for inferring other mappings between related concepts.

P2P ontology mapping[32]: This work[32] proposes the framework which allows agents to interact with other agents efficiently based on the dynamic mapping of only the portion of ontologies relevant to the interaction. The framework executes three steps: 1) Generates the hypotheses. 2) Filters the hypotheses. 3) Selects the best hypothesis.

3.3 Ontology mapping (matching) in ontology merging and alignment

This category allows a single coherent merged ontology to be created through an ontology merging process. It also creates links between local ontologies while they remain separate during the ontology alignment process. Mappings do not exist between a single coherent merged ontology and local ontologies, but rather between local ontologies to be merged or aligned. Defining a mapping between local ontologies to be merged or aligned is the first step in the ontology merging or alignment process. This mapping identifies similarities and conflicts between local ontologies to be merged or aligned.

3.3.1 Strength and drawbacks

This mapping applies to ontologies over the same or overlapping domain. Finding mapping is a part of other applications such as ontology merging or alignment. This might be fairly obvious and more interesting in a large ontology.[14,11]

3.3.2 Application domains

The growing usage of ontologies or the distributed nature of ontology development has led to a large number of ontologies which have the same or overlapping domains.[15,17] These should be merged or aligned to be reused.[15] Many applications such as standard search, e-commerce, government intelligence, medicine, etc., have large-scale ontologies and require the reuse of ontology merging processes.[11]

3.3.3 Tools, systems, and related work

SMART[18]: SMART is a semi-automatic ontology merging and alignment tool. It looks for linguistically similar class names through class-name matches, creates a list of initial linguistic similarity (synonym, shared substring, common suffix, and common prefix) based on class-name similarity,

studies the structures of relation in merged concepts, and matches slot names and slot value types. It makes suggestions for users, checks for conflicts, and provides solutions to these conflicts.

PROMPT[15]: PROMPT is a semi-automatic ontology merging and alignment tool. It begins with the linguistic-similarity matches for the initial comparison, but generates a list of suggestions for the user based on linguistic and structural knowledge and then points the user to possible effects of these changes.

OntoMorph[16]: OntoMorph provides a powerful rule language for specifying mappings, and facilitates ontology merging and the rapid generation of knowledge-base translators. It combines two powerful mechanisms for knowledge-base transformations such as syntactic rewriting and semantic rewriting. Syntactic rewriting is done through pattern-directed rewrite rules for sentence-level transformation based on pattern matching. Semantic rewriting is done through semantic models and logical inference.

HICAL[19] (Hierarchical Concept Alignment system): HICAL provides concept hierarchy management for ontology merging/alignment (one concept hierarchy is aligned with another concept in another concept hierarchy), uses a machine-learning method for aligning multiple concept hierarchies, and exploits the data instances in the overlap between the two taxonomies to infer mappings. It uses hierarchies for categorization and syntactical information, not similarity between words, so that it is capable of categorizing different words under the same concept.

Anchor-PROMPT[20]: Anchor-PROMPT takes a set of anchors (pairs of related terms) from the source ontologies and traverses the paths between the anchors in the source ontologies. It compares the terms along these paths to identify similar terms and generates a set of new pairs of semantically similar terms.

CMS[23] (CROSI Mapping System): CMS is an ontology alignment system. It is a structure matching system on the rich semantics of the OWL constructs. Its modular architecture allows the system to consult external linguistic resources and consists of feature generation, feature selection, multi-strategy similarity aggregator, and similarity evaluator.

FCA-Merge[17]: FCA-Merge is a method for ontology merging based on Ganter and Wille's formal concept analysis[28], lattice exploration, and instances of ontologies to be merged. The overall process of ontology merging consists of three steps: 1) instance extraction and generation of the formal context for each ontology, 2) the computation of the pruned concept lattice by algorithm TITANIC[29], and 3) the non-automatic generation of the merged ontology with human interaction based on the concept lattice.

CHIMAERA[30]: CHIMAERA is an interactive ontology merging tool based on the Ontolingual ontology editor. It makes users affect merging process at any point during merge process, analyzes ontologies to be merged, and if linguistic matches are found, the merge is processed automatically, otherwise, further action can be made by the use. It uses subclass and super class relationship.

3.4 A Comparison of ontology mapping tools or systems

A specific unified framework does not exist for comparison of ontology mapping tools[2], nor may direct comparison of ontology mapping tools be possible.[10] But a set of evaluation criteria to compare ontology mapping tools is proposed[10] and some of systems about ontology mapping are compared.[8] See Table 1 for a summary of ontology mapping tools.

4. Conclusion

This paper has presented a broad scope of ontology mapping, mapping categories and characteristics, and surveyed ontology mapping tools, systems, and related work based on ontology mapping categories as follows: a mapping between an integrated global ontology and local ontologies, a mapping between local ontologies, and a mapping on ontology merging and alignment. The different roles of these three ontology mapping categories were also identified. Techniques for a mapping between local ontologies have not been widely used for a mapping between a global ontology and local ontologies for two reasons. First, mapping between a global ontology and local ontolgies is done in the process of ontology integration or when a global ontology exists.[3, 4, 7] Second, some techniques for a mapping between local ontolgies are aimed at distributed ontologies on the Semantic Web, ontologies which have mutually inconsistent concepts or requirements of a more dynamic or flexible form of mapping.[1, 6, 8, 9, 12, 22, 32]

Further research is needed to improve methods of constructing an integrated global ontology, utilizing the mapping techniques for local ontologies in order to map between an integrated global ontology and local ontologies. In addition, research about the usage or roles of ontology mapping in different application domains should be performed. Research aimed at developing sufficiently applicable mapping techniques between local ontologies for the same or overlapping domain will improve ontology merge and alignment processes. In order to find an accurate ontology mapping, accurate similarity measurements between source ontology entities and target ontology entities should be considered. Techniques for complex ontology mappings between

ontologies and discovering more constraints in ontologies should be also investigated.

	MOMIS	LSD	CTXMATCH	GLUE	MAFRA	LOM	ONION	PROMPT	FCA-Merge
Input	Data model	Source schemas & their instances	Concepts in concept hierarchy	Two taxonomies with their data instances in ontologies	Two ontologies	Two lists of terms from two ontologies	Terms in two ontologies	Two input ontologies	Two input ontologies and a set of documents of concepts in ontologies
Output	An integrated global ontology (GVV)	pairs of related terms between a global and local schema	Semantic relation between concepts	A set of pairs of similar concepts	Mappings of two ontologies by the Semantic bridge ontology	A list of matched pairs of terms with score ranking similarity	Sets of Articulation rules between two ontologies	A merged ontology	A merged ontology
User interaction	The designer involves in schema annotation & sets a threshold for integration clusters for generating a GVV	The user provides mappings for training source & feedback on the proposed mappings.	No (CTXMATCH is an algorithm.)	User-defined mappings for training data , similarity measure, setting up the learner weight, and analyzing system's match suggestion	The domain expert interface with the similarity and semantic bridging modules and it has graphical user interface	It requires human validation at the end of the process.	A human expert chooses or deletes or modifies suggested matches using a GUI tools	The user accepts, Rejects , or adjusts system's suggestions.	Generating a merged ontology requires human interaction of the domain expert with background knowledge
Mapping strategy or algorithm	Name equality: Synonyms hyponyms Matching of clustering	Multi-strategy Learning approach : (machine Learning technique)	Logical deduction	Multi-strategy learning approach : (machine learning technique)	Semantic bridge	Lexical similarity whole term, word constituent, synset, and type matching	Linguistic matcher, Structure-, inference-based heuristics	Heuristic-based analyzer	Linguistic analysis & TITANIC algorithm for computation for pruned concept lattice
Structured knowledge	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes
Instance-based knowledge	No	Yes	No	Yes	Yes	No	No	No	Yes
Lexical knowledge	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
domain knowledge	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes

Table 1 A summary of ontology mapping tools

5. References

1. Alexander Maedche, Boris Motik, Ljiljana Stojanovic, Rudi Studer, and Raphael Volz, "*Ontologies for Enterprise Knowledge Management*", IEEE Intelligent Systems, 2003.

2. Yannis Kalfoglou, Marco Schorelmmer, *"Ontology Mapping: The State of the Art"*, The Knowledge Engineering Review, Vol. 18:1, 1-31, 2003.

3. AnHai Doan, Pedro Domingos, Alon Halevy, "*Learning to Match the Schemas of Data Sources: A Multistrategy Approach"*, Machine Learning, 50 (3): 279-301, March 2003.

4. Domenico Beneventano, Sonia Bergamaschi**,** Francesco Guerra, Maurizio, "*Synthesizing an Integrated Ontology"*, IEEE Internet Computing, September • October 2003.

5. Xiaomeng Su, "*Semantic Enrichment for Ontology Mapping"* PhD thesis Dept. of Computer and Information Science, Norwegian University of Science and Technology.

6. Paolo Bouquet, Fausto Giunchiglia, Frank van Harmelen, Luciano Serafini, Heiner Stuckenschmidt, "*C-OWL: Contextualizing Ontologies*", ISWC 2003, LNCS 2870, pp.164-179, 2003.

7. Calvanese, D, De Giacomo, G and Lenzerini, M, 2001a, "*A Framework for Ontology Integration*" Proceedings of the 1st International Semantic Web Working Symposium (SWWS) 303–317.

8. Paolo Bouquet, Luciano Serafini, Stefano Zanobini, "*Semantic Coordination: A New Approach and an Application"*, ISWC 2003, LNCS 2870, pp.130-145, 2003.

9. AnHai Doan, Jayant Madhavan, Pedro Domingos, Alon Halevy, "*Learning to Map between Ontologies on the Semantic Web*", VLDB Journal, Special Issue on the Semantic Web, 2003.

10. N. F. Noy and M.A. Musen, "*Evaluating Ontology*

Mapping Tool: Requirement and Experience", Proceedings of the Workshop on Evaluation of Ontology Tools at EKAW'02 (EOEN2002), Siguenza, Spain, 2002.
11. Deborah L. McGuinness, Richard Fikes, James Rice, Steve Wilder, *"An Environment for Merging and Testing Large Ontologies"*, Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR200), Breckenridge, CO, April 12-15, 2000.
12. Nuno Silva, Joao Rocha, *"MAFRA – An Ontology Mapping FRAmework for the Semantic Web"*, Proceedings of the 6th International Conference on Business information Systems; UCCS, Colorado Springs, CO, May 2003.
13. Mitra, P and Wiederhold, G, "*Resolving Terminological Heterogeneity in Ontologies*", Proceedings of the ECAI'02 workshop on Ontologies and Semantic Interoperability, 2002.
14. Madhavan, J, Bernstein, PA, Domingos, P and Halevy, A, , "*Representing and reasoning about mappings between domain models*", Proceedings of the 18th National Conference on Artificial Intelligence (AAAI'02), 2002.
15. N. Noy and M. Musen, *"PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment.*" Proceedings of the National Conference on Artificial Intelligence (AAAI), 2000.
16. H. Chalupsky. "*Ontomorph: A Translation System for Symbolic Knowledge*", Principles of Knowledge Representation and Reasoning, 2000.
17. Gerd Stumme, Alexander Maedche, *"FCA-Merge: Bottom-Up Merging of Ontologies"*, In proceeding of the International Joint Conference on Artificial Intelligence IJCA101, Seattle, USA, 2001.
18. Natalya Fridman Noy and Mark A. Musen, *"Smart: Automated Support for Oontology Merging and Alignment"*, Proceedings of the Twelfth Banff Workshop on Knowledge Acquistion, Modeling, and Management, Banff Algeberta, 1999.
19. R. Ichise, H. Takeda, and S. Honiden. "*Rule Induction for Concept Hierarchy Alignment*", Proceedings of the Workshop on Ontology Learning at the 17th International Joint Conference on Artificial Intelligence (IJCAI), 2001.
20. N. Noy and M. Musen, *"Anchor-PROMPT: Using Non-Local Context for Semantic Matching*", Proceedings of the Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI), 2001.
21. John Li, *"LOM: A Lexicon-based Ontology Mapping Tool"*, Proceedings of the Performance Metrics for Intelligent Systems (PerMIS. '04), 2004.
22. Marc Ehrig, Steffen Staab, *"QOM – Quick Ontology Mapping"*, GI Jahrestagung (1), 2004.
23. Yannis Kalfoglou, Bo Hu, *"CROSI Mapping System (CMS) Results of the 2005 Ontology Alignment Contest"*, K-CAP Integrating Ontologies Workshop 2005, Banff, Alberta, Canada, 2005.
24. Helena Sofia Pinto, Joao P. Martins, *"A Methodology for Ontology Integration*", Proceedings of the International Conference on Knowledge Capture, Technical papers, ACM Press, pp. 131-138, 2001.
25. Nuno Silva and Joao Rocha, *"Ontology Mapping for Interoperability in Semantic Web"*, Proceedings of the IADIS International Conference WWW/Internet 2003 (ICWI'2003). Algarve, Portugal; November 2003.
26. H. Sofia Pinto, A. Gomez-Perez, J. P. Martins, *"Some Issues on Ontology Integration"*, In Proc. of IJCAI99's Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends, 1999.
27. Studer R, Benjamins VR, Fensel D, *"Knowledge Engineering: Principles and Methods"*, IEEE Transactions on Data and Knowledge Engineering, 25(1-2): 161- 199, 1998.
28. Ganter B., Wille R., *"Formal Concept Analysis: Mathematical Foundaations Springer"*, 1999.
29. G. Stumme, R. Taouil, Y. Bastide, N. Pasquiter, L. Lakhal, "Fast computation of concept lattices using data mining techniques", Proc. KRDB '00, http: //sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS 129-139, 2000.
30. D. McGuinness, R. Fikes, J. Rice, and S. Wilder, "*The Chimaera Ontology Environment*", In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI), 2000.
31. Prasenjit Mitra, Natasha F. Noy, Anju Jaiswals, *"OMEN: A Probabilistic Ontology Mapping Tool"*, International Semantic Web Conference 2005: 537-547.
32. Paolo Besana, Dave Robertson, and Michael Rovatsos, *"Exploiting interaction contexts in P2P ontology mapping"*, P2PKM2005.
33. Boddy, M., *"Anytime problem solving using dynamic programming"*, In proceedings of the Ninth National Conference on Artificial Intelligence, Anaheim, California, Shaker Verlag (1991) 738-743.
34. D. Fensel, *"Ontologies: Silver Bullet for knowledge management and Electronic Commerce"*, Springer_Verlag, 2001.
35. Natalya F. Noy, Michel Klein, *"Ontology Evolution: Not the same as Schema Evolution"*, In: Knowledge and Information Systems, 6(4): 428-440, July, 2004.
36. C. Batini and M. Lenzerni. *"A comparative analysis of methodologies for database schema integration"*, ACM Computer Surveys, 18(4), 1986.
37. AnHai Doan, Alon Y. Halevy, *"Semantic Integration Research in the Database Community: A Brief Survey"*, AI Magazine, Volume 26, Mar. 2005.
38. Madhavan J., Bernstein P., Doan A., and Halevy A., *"Corpus based schema matching"*, In proc. of The 18th IEEE Int. Conf. on Data Engineering, 2005.

The Database Research Group at the Max-Planck Institute for Informatics

Gerhard Weikum
Max-Planck Institute for Informatics
Stuhlsatzenhausweg 85
D-66123 Saarbruecken, Germany
weikum@mpi-inf.mpg.de

1. INTRODUCTION

The Max-Planck Institute for Informatics (MPI-INF) is one of 80 institutes of the Max-Planck Society, Germany's premier scientific organization for foundational research with numerous Nobel prizes in natural sciences and medicine. MPI-INF hosts about 150 researchers (including graduate students) and comprises 5 research groups on algorithms and complexity, programming logics, computational biology and applied algorithmics, computer graphics, and databases and information systems (DBIS). This report gives an overview of the DBIS group's mission and ongoing research.

2. VISION AND RESEARCH DIRECTIONS

The research of the DBIS group pursues two major directions:

1. intelligent organization and search of semistructured information, in intranets, digital libraries, and on the Web;
2. architecture and strategies for self-organizing distributed information systems, particularly, peer-to-peer systems.

2.1 Intelligent Organization and Search of Information

The age of information explosion poses tremendous challenges regarding the intelligent organization of data and the effective search of relevant information in business and industry (e.g., market analyses, logistic chains), society (e.g., health care), and virtually all sciences that are more and more data-driven (e.g., gene expression data analyses and other areas of bioinformatics). The problems arise in intranets of large organizations, in federations of digital libraries and other information sources, and in the most humongous and amorphous of all data collections, the World Wide Web and its underlying numerous databases that reside behind portal pages. The Web bears the potential of being the world's largest encyclopedia and knowledge base, that would be of great value to all kinds of "knowledge workers" from students to Nobel laureates, but we are very far from being able to exploit this potential.

Search-engine technologies provide support for organizing and querying information; for simple mass-user queries that aim to find popular Web pages on pop stars, soccer clubs, or the latest Hollywood movies, existing engines like Google are probably the best solution. But for advanced information demands search engines all too often require excessive manual preprocessing, such as manually classifying documents into a taxonomy for a good Web portal, or manual postprocessing such as browsing through large result lists with too many irrelevant items or surfing in the vicinity of promising but not truly satisfactory approximate matches. The following are example queries where current Web and intranet/enterprise search engines fall short:

Q1: Which professors from Saarbruecken in Germany teach information retrieval and participate in EU projects?

Q2: Which drama has a scene in which a woman makes a prophecy to a Scottish nobleman that he will become king?

Q3: Who was the French woman that I met in a program committee meeting where Paolo Atzeni was the PC chair?

Why are these queries difficult (too difficult for current Web search engines, unless one invests a huge amount of time to manually explore large result lists with mostly irrelevant and some mediocre matches)? For Q1 no single Web site is a good match; rather one has to look at several pages together within some bounded *structural context*: the homepage of a professor with his address, a page with course information linked to by the homepage, and a research project page that is a few hyperlinks away from the homepage.

Q2 cannot be easily answered because a good match does not necessarily contain the keywords "woman", "prophecy", "nobleman", etc., but may rather say something like "Third witch: All hail, Macbeth, thou shalt be king hereafter!" and the same document may contain the text "All hail, Macbeth! hail to thee, thane of Glamis!". So this query requires some *background knowledge* to recognize that a witch is usually female in the literature, "shalt be" refers to a prophecy, and "thane" is a title for a Scottish nobleman.

Q3 combines the difficulties of Q1 and Q2 as it requires background knowledge but also needs to put together bits and pieces from different information sources, including semistructured desktop data: PC meetings that I attended

according to my electronic calendar, conferences on which I served on the PC found in my email archive, PC members listed in electronic proceedings, and detailed information found on researchers' homepages. And after having identified a candidate like Sophie Cluet from Paris, one needs to infer that Sophie is a typical female first name and that Paris most likely denotes the capital of France rather than the 500-inhabitants town of Paris, Texas, which became known through a movie.

Answering such queries with high precision, despite high diversity and noise in the underlying data, calls for a new kind of *"semantic search" engine* that combines concepts and techniques from database and information-retrieval (IR) systems. A promising starting point is *XML IR* [1], especially when combined with background knowledge in the form of ontologies and thesauri. We will outline our approach along these lines in Subsection 3.1.

The envisioned "semantic search" should not only address the querying itself, but also aim at *intelligent organization of information*. Data with more explicit structure, semantic annotations, and clean data items (e.g., reconciled names of persons, organizations, etc.) becomes easier to search with high precision. But the Web and also federations of digital libraries and other data sources are highly diverse in terms of structure, annotations, and data quality (e.g., authority, resolution, completeness, freshness, etc.). Therefore, we also pursue ways of automatically annotating and structuring information, using techniques from natural language processing (NLP) and statistical machine learning. We will outline our approaches in this area in Subsection 3.3.

2.2 Self-organizing Distributed Information Systems

Even when we focus on scientific information alone and leave out business and entertainment data, new information is produced and compiled world-wide in a highly distributed manner, in federations of digital libraries or e-science repositories and, of the course, on the Web with millions of scholars and students. Thus, it is natural to pursue a completely decentralized peer-to-peer (P2P) architecture [30] for managing this information explosion, intelligently organizing the information, and efficiently searching it.

The P2P approach bears the potential of overcoming the shortcomings of today's Web search engine technology and successfully tackling the kinds of killer queries mentioned in Subsection 2.1. In our architecture every peer, e.g., the home PC of a scientist or student, has a full-fledged search engine that indexes a small portion of the Web, according to the interest profile of the user. Such an architecture has four major advantages over a centralized server farm:

1. As the data volume and the query load per peer are much lighter, the peer's search engine can employ much more advanced techniques for concept-based rather than keyword-based search, leveraging background knowledge in the form of thesauri and ontologies and powerful mathematical and linguistic techniques such as spectral analysis and named entity recognition.

2. Peers can collaborate for finding better answers to difficult queries: if one peer does not have a good result locally it can contact a small number of judiciously chosen peers who are considered "knowledgeable" on the query topic. This approach should often be able to exploit the small-world phenomenon on the Web: knowledgable peers are only a short distance away.

3. A P2P system can gather and analyze bookmarks, query histories, user click streams, and other data about user and community behavior; the implicit and explicit assessments and recommendations derived from this input can be leveraged for better search results. In contrast to a central server, the P2P approach provides each user with direct, fine-grained control over which aspects of her behavior may be collected and forwarded to other peers.

4. A politically important issue is that a P2P search engine is less susceptible to manipulation, censorship, and the bias induced by purchased advertisements.

Thus, a P2P approach to information search could pave the way towards a *"social search"* super-engine that leverages the collaborative "wisdom of crowds" [14]. Challenges that we face towards this vision are scalability and the desire to make the P2P network self-organizing and resilient to high dynamics (failures and churn) and possible manipulation (cheating, spam, etc.). We will outline our approach to P2P search in Subsection 3.2. Again, search needs to be complemented by powerful Web mining, which will be discussed in Subsection 3.3.

3. ONGOING PROJECTS

3.1 XML Ranked Retrieval: TopX and Sphere Search

Non-schematic XML data that comes from many different sources and inevitably exhibits heterogeneous structures and annotations (i.e., XML tags) cannot be adequately searched using database query languages like XPath or XQuery. Often, queries either return too many or too few results. Rather the ranked-retrieval paradigm is called for, with relaxable search conditions, various forms of similarity predicates on tags and contents, and quantitative relevance scoring.

TopX [35] is a search engine for ranked retrieval of XML data, developed at MPI-INF. TopX supports a probabilistic-IR scoring model for full-text content conditions and tag-term combinations, path conditions for all XPath axes as exact or relaxable constraints, and ontology-based relaxation of terms and tag names as similarity conditions for ranked retrieval.

While much of the TopX functionality was already supported in our earlier work on the XXL system [32, 33], TopX has an improved scoring model for better precision and recall, and it is much more efficient and scalable. TopX has been stress-tested and experimentally evaluated on a variety of datasets including the TREC Terabyte benchmark, the INEX XML information retrieval benchmark, and an XML version of the Wikipedia encyclopedia. For the INEX 2006 benchmark [17], TopX serves as the official

reference engine for topic development and several benchmarking tasks. For good performance, TopX employs various novel techniques: carefully designed index structures [24, 35], probabilistic models as efficient score predictors [34, 35], judicious scheduling of index accesses [35, 3], and the incremental merging of index lists for on-demand, self-tuning query expansion [36].

For top-k similarity queries we have extended the TA family of threshold algorithms [13]. Large disk-resident index lists for content terms and tag names strongly suggest using the TA variant with sorted access only for disk I/O efficiency. This method maintains a priority queue of score intervals for candidate items and performs a conservative threshold test based on upper bounds of candidate scores. Our new method uses a probabilistic threshold test for fast approximation of top-k results. It is based on score predictors that compute convolutions of score distributions in the yet to be scanned tails of index lists. This method provides an order of magnitude improvement in run-time at high and probabilistically controllable levels of precision and recall relative to the TA baseline [34, 35]. Our recent techniques for judicious scheduling of block-level sequential and item-level random accesses yields an additional performance gain by another factor of five [3].

For enhanced quality of search results, relevance feedback is a well-known IR technique that expands keyword queries from the content of elements marked as relevant by the user. We have developed novel feedback techniques that expand a keyword query into a possibly complex content-and-structure query that specifies additional XPath conditions on the structure of the desired results [25]. This way the user is largely relieved from understanding the subtleties of XPath and the XML structure of the underlying, possibly highly heterogeneous, XML documents. As an example, consider a query about the life of the physicist Max Planck on a richly tagged XML version of Wikipedia. As a keyword query, the user may formulate this request as "life physicist Max Planck". Unfortunately, the results may be dominated by articles about Max Planck institutes, for example, in the area of life sciences. By providing the feedback that elements on persons and their biographies and the topic physics in mediocre result documents are preferred over completely irrelevant results, the engine would ideally be able to automatically generate a content-and-structure query such as

```
//article[.//person ftcontains(‘‘Max Planck’’)]
        [.//category ftcontains(‘‘physicist’’)].
```

This XPath Full-Text query would yield much more precise results.

As XML data is still rare on the surface Web and even structured Deep-Web sources do often expose only HTML pages, we have started working on converting Web data into XML format. This involves identifying and tagging named entities like persons, organizations, locations, time-points, and time periods. To this end, we are using open-source tools like ANNIE from the University of Sheffield [12] and MinorThird from CMU [11], resulting in a richly (but heuristically) tagged corpus that unifies XML and HTML and other Web documents by casting all data into XML. As Web pages are often highly inter-linked, the resulting corpus is no longer a repository of XML trees, but includes many XLinks (created from href hyperlinks) leading to an arbitrary XML graph. We have developed the SphereSearch prototype engine [15] that can search such graph data and rank search results, using a query language that is simpler than XPath Full-Text but much more powerful than keyword-based Web search. The result of a query is a subgraph spanned by nodes that approximately match and have high scores for the query's elementary conditions. Finding the top-k results involves an approximation to the Steiner tree problem, based on mimimum spanning trees on a precomputed connection graph, using a top-k threshold algorithm.

TopX and SphereSearch are incomparable in terms of query expressiveness. TopX is generally more efficient, but currently limited to XML trees. Our future research will aim at extending TopX to include the richer graph-based search capabilities of SphereSearch in a highly efficient and scalable manner.

3.2 Peer-to-Peer Search: Minerva

For P2P search over Web data with ranked retrieval, we are developing the Minerva system [5].[1] Each peer has a full-fledged Web search engine, including a crawler and an index manager. The crawler may be thematically focused or crawl results may be postprocessed so that the local index contents reflects the corresponding user's interest profile. For collaborative search, the peers are connected by by an overlay network based on a distributed hash table (DHT). The DHT also forms the basis of a conceptually global but physically decentralized and scalable directory that contains metadata and statistics about the peers' contents and quality. Note that, for scalability, the directory is not designed as a page-granularity global Web index, but is limited in size to the number of indexed features (e.g., keywords or topics) times the number of peers. This avoids the pitfalls outlined in [19]. Also note that the pursued P2P Web search includes ranked retrieval and is thus fundamentally much more difficult than Gnutella-style file sharing or simple key lookups via DHTs.

With a user's highly specialized and personalized "power search engine" most queries should be executed locally, but once in a while the user may not be satisfied with the local results and would then want to contact other peers. This is the *query routing* (or peer-selection) problem, the cornerstone of the P2P search engine. Although the problem is related to earlier work on metasearch engines and distributed information retrieval [21], the P2P setting is much more challenging because of larger scale and high dynamics. A "good" peer to which the user's query should be forwarded would have thematically relevant index contents, which could be measured by statistical notions of similarity between peers. On the other hand, each additional target peer for a query should yield novel results that are not yet provided by previously selected peers or even the local index of the query initiator itself. To this end, we have developed an overlap-aware query routing strategy that aims to optimize a weighted combination of search result quality and novelty [4, 6]. As the query routing decision made for execution planning is in the critical path of user-perceived response time, fast estimation of quality-novelty measures is crucial. We have developed

[1]Minerva is the Roman goddess of science, wisdom, and learning, and is also the icon of the Max Planck Society.

new methods for this purpose, utilizing compact synopses like hash sketches and min-wise independent permutations in combination with the underlying DHT.

For the actual top-k query processing in the P2P network, it is sometimes necessary to aggregate index entries from multiple peers, combining their local scores into a global quality measure. The KLEE algorithm for distributed top-k queries [22] aims to minimize network latency, network bandwidth consumption, and the local CPU and disk IO cost of the participating peers. KLEE proceeds in three or, optionally, four phases, driven by the query originator as a per-query coordinator.[2] These phases serve to 1) determine an initial set of top-k candidates and obtain score-distribution histograms and Bloom-filter summaries from peers, 2) estimating the cost/benefit ratio of requesting additional synopses, 3) optionally obtaining this extra information, and 4) obtaining missing scores for the remaining candidates and computing the total scores for the final top-k result. KLEE has been implemented in the Minerva testbed, and has consistently outperformed various competitors on several real-life datasets and query benchmarks.

The above outlines of query routing and query processing in a P2P network show that distributed management of statistical information about peers and their data collections is a key issue. This involves efficient gathering and dissemination of statistics as well as estimations for specific purposes. A global measure of particular interest (for query routing and query result merging) is the document frequency (df) of a keyword, i.e., the total number of distinct documents in the entire network that contain the keyword. Estimating this number is difficult because of duplicate documents at different peers. We have developed a highly efficient and accurate method for this problem, by combining hash sketches and the DHT-based overlay network [8]. Another network-wide statistical estimation problem is to efficiently determine pairs of highly correlated or anti-correlated keywords, either in queries or in the data. In [7] we have developed a technique that utilizes the DHT-based infrastructure for an efficient solution, which can piggyback all necessary message exchanges on the network traffic that is needed for standard query routing and execution anyway. Awareness of keyword correlations is useful for more effective query routing decisions.

As mentioned in Subsection 2.2, a P2P network is a natural habitat for "social search" that leverages community assessments. A simple form of such community input are link analysis methods like Google's PageRank or Kleinberg's HITS. But these are centralized algorithms with very high memory demand, and their distributed variants assume that the underlying Web graph can be nicely partitioned among sites. In contrast, a P2P system like Minerva emphasizes the autonomy of peers that can crawl Web fragments and gather their own local content at their discretion. JXP [23] is a new algorithm for dynamically computing, in a decentralized P2P manner, global authority scores when the Web graph is spread across many autonomous peers. In this setting, the peers' graph fragments may overlap arbitrarily, and peers are (a priori) unaware of other peers' fragments. With JXP, each peer computes the authority scores of the pages that it has in its local index, by locally running the standard PageRank algorithm. A page may be known and indexed by multiple peers, and these may have different scores for that same page. A peer gradually increases its knowledge about the rest of the network by meeting with other, randomly chosen, peers and exchanging information, and then recomputes the PageRank scores for its pages of interest. The local computations are very space-efficient (as they require only the local graph and the authority-score vector), and fast (as they operate on much smaller graph fragments than a server-side global PageRank computation). We have proven, using the theory of Markov-chain aggregation/disaggregation, that the JXP scores do indeed converge to the same values as a global PageRank computation on the full Web graph [23].

[2]Klee is an expressionistic painter, but also the German word for clover, which has usually three leaves but infrequently occurs with four leaves. The latter is traditionally viewed as a symbol of good luck.

3.3 Web Mining

The duality between better organization of information and more effective search has motivated us to embark also on various aspects of Web mining, so as to provide a richer basis for better search capabilities. The following subsections briefly discuss the work along these lines.

3.3.1 Query Logs and Click Streams

Information about user behavior is a rich source to build on. This includes relatively static properties like bookmarks or embedded hyperlinks pointing to high-quality Web pages, but also dynamic properties inferred from query logs and click streams. For example, suppose a user clicks on a specific subset of the top 10 results returned by a search engine for a query with several keywords, based on having seen the summaries of these pages. This implicit form of relevance feedback establishes a strong correlation between the query and the clicked-on pages. When a user does not click on any of the top 10 results for a given query and rather chooses to rephrase the query using different keywords, this may be interpreted as negative feedback on the relevance or quality of the initial results. Exploiting this kind of user-behavior information can help to improve search result quality for individuals as well as entire communities by adding more "cognitive" elements to the search engine.

Our approach is based on a Markov-chain model with queries as additional nodes, additional edges that capture query refinements and result clicks, and corresponding transition probabilities. Similarly to Google's PageRank, the computation of stationary visiting probabilities yields the behavior-aware authority ranking. Our original model could not express negative feedback, as probabilities are non-negative and L1-normalized. To capture and exploit also negative assessment such as assigning trust levels to Web pages (e.g., marking a Web page as spam, low-quality, out-of-date, or untrusted), we are pursueing several extended approaches, one of which is based on a Markov reward model where the assessment part is uncoupled from the random walk in the extended Web graph. A page receives a specific lump reward each time a transition is made to it, and this reward depends on the transition's source and target and will be derived from the query-log and click-stream information as well as explicit page assessments. The reward itself can be positive or negative.

The most interesting measure in such reward models is the expected earned reward per time-step. Preliminary results with this approach are presented in [20].

3.3.2 Web Evolution

Today, virtually all Web repositories, including digital libraries and the major Web search engines, capture only current information. But the history of the Web, its lifetime over the last 15 years and many years to come, is an even richer source of information and latent knowledge. It captures the evolution of digitally born content and also reflects the near-term history of our society, economy, and science. Web archiving is done by the Internet Archive, with a current corpus of more than 2 Petabytes and a Terabyte of daily growth, and, to a smaller extent, some national libraries in Europe. These archives have tremendous latent value for scholars, journalists, and other professional analysts who want to study sociological, political, media usage, or business trends, and for many other applications such as issues of intellectual property rights. However, they provide only very limited ways of searching timelines and snapshots of historical information.

We are working on Web search and ranking of authoritative pages with a user-adjustable time focus, supporting both snapshots and timelines. Our notions of T-Rank and BuzzRank [9, 10] provide time-aware generalizations of the PageRank importance measure. As an example for the benefits obtained from time-aware ranking, consider looking for a database publication of the year 1993 with a strong trend of increasing popularity. When you apply static PageRank to Web sources (including portals such as DBLP), you obtain Codd's seminal paper as the best result, whereas BuzzRank identifies the Association-Rule-Mining paper by Agrawal, Imielinski, and Swami as the best emerging authority. As the underlying link matrices for these computations are huge, one of the challenges that we are addressing is to implement these new kinds of extended link analyses in an efficient and scalable manner. To this end, we are investigating compact representations for time-series of link graphs, so as to reduce both the space and time complexity of time-aware ranking and time-travel queries over Web archives.

3.3.3 Ontology Learning

Several of our projects, especially the TopX engine for XML IR, make use of ontologies and thesauri. The latter can be constructed, to some extent, from hand-crafted knowledge such as WordNet, but a larger-scale and self-maintaining, promising approach is to automatically learn the concepts and relations for an ontology directly from rich text corpora such as Wikipedia and other Internet sources. This requires a combination of linguistic analysis, pattern matching, and statistical learning to identify, for example, person names or locations and to extract instances of binary relations such as `located-at (city, river)`, `born-in (person, place)`, `plays-instrument (person, instrument)`, or the generic `is-instance-of (entity, concept)`.

For finding and extracting information on binary relations we pursue a novel approach based on a link-grammar representation of natural-language sentences and an SVM-based statistical learner for determining robust, generalizable linguistic patterns. This approach is implemented in the LEILA prototype system [31]. The method is almost unsupervised by starting with merely a small set of user-provided positive examples such as *(Paris, Seine), (Calcutta, Ganges), (London, Thames)* and either explicit or token-based negative examples such as *(New York, Mississippi)* or *(<proper noun>, <number>*. The system then automatically finds candidate patterns such as *<river> flows through <city>*, *<city> is located on the banks of the <river>*, or *<city> not only offers many cultural attractions, but visitors can also enjoy tubing or swimming in the <river>*. These patterns cannot be directly applied as they do not generalize well and would lead to many false positives. Instead, we run the CMU link-grammar parser [18] on the individual sentences, compute a characteristic feature vector from the resulting graph representation of each sentence, and feed these vectors into an SVM classifier. For higher recall, we also perform anaphora resolution across neighboring sentences. The method outperforms competitors in terms of the F1 measure (i.e., the harmonic mean of precision and recall). We are currently working on further extensions and better scalability.

3.3.4 Document Classification

Automatic classification of text documents is an important building block for many forms of intelligent data organization, for example: assigning Web pages to topics of a hierarchical directory, filtering news feeds or thematic subscriptions in digital libraries, steering a focused crawler towards Web regions that are more likely to contain pages that fall into the user's specific interest profile. Classifiers are models over high-dimensional feature spaces that are derived from supervised learning, with positive and negative examples as training input. The underlying methods for statistical learning such as SVM or Bayesian methods have become fairly mature, and the practical bottleneck is typically the scarceness of training data, because compiling training samples is a laborious and time-consuming human activity. Our research is primarily aiming at overcoming this training bottleneck and improving the accuracy and robustness of text classifiers.

We have developed a family of adjustable metamethods [26, 27, 28] based on ensemble learning that can be tuned towards the specific goals of the classifier: when precision and accuracy are critical we can prioritize the mutual agreement of multiple classifiers, but when recall is more important we can relax the settings of the meta-classifier. A particular application of these methods is our focused crawler BINGO! [29], which can start with a small set of training samples and seed URLs and then automatically finds characteristic "archetype" pages for dynamic and automated re-training. In this semisupervised-learning context, it is crucial to minimize the classification error, and restrictive metamethods can be effectively tuned towards this goal.

In a second line of research we investigate richer feature models and contextual features of training samples so as to improve classification accuracy with relatively small training sets. One approach is to consider neighbors of text documents in environments with many cross-references such as Web links (but covering also settings such as book, music, or blog recommendations or citation graphs for publications). The graph-based classifier presented in [2] builds on the theory of Markov Random Fields, and addi-

tionally develops new techniques for enhanced robustness. This method outperforms purely text-based classifiers and also all previously proposed link-aware classifiers. Finally, we investigate new models that map text features onto semantic concepts in an ontology and leverage the concepts as additional information. In [16] we have developed a generative probabilistic model that uses concepts as a latent-variable layer and an efficient EM-based parameter-estimation procedure. In contrast to previous work on latent semantic models, concepts are explicit and directly interpretable by humans and greatly simplify the model-selection problem of choosing the right number of latent dimensions. In combination with a transductive learning procedure that leverages the richer feature space of unlabeled background corpora, we have built a highly effective classifier that excels especially when training data is very scarce.

4. ACKNOWLEDGEMENTS

All members of the DBIS group at MPI-INF have been greatly contributing to our research results and ongoing projects. As of June 2006, the group has the following scientific staff and graduate students (born and raised in 7 different countries): Ralitsa Angelova, Srikanta Bedathur, Matthias Bender, Klaus Berberich, Andreas Broschart, Tom Crecelius, Jens Graupmann, Georgiana Ifrim, Gjergji Kasneci, Julia Luxenburger, Sebastian Michel, Thomas Neumann, Hanglin Pan, Josiane Parreira, Maya Ramanath, Ralf Schenkel, Stefan Siersdorfer, Fabian Suchanek, Martin Theobald, Gerhard Weikum, Christian Zimmer. In addition, several visitors and external collaborators have contributed, most notably, Peter Triantafillou and Michalis Vazirgiannis. Our projects have various funding sources. We are grateful to the German Science Foundation (DFG) and the Commission of the European Union (EU) for supporting the projects CLASSIX (DFG), DELIS (EU), and the Network of Excellence DELOS (EU).

5. REFERENCES

[1] S. Amer-Yahia et al.: Report on the DB/IR Panel at SIGMOD 2005, SIGMOD Record 34(4): 71-74, 2005.

[2] R. Angelova, G. Weikum: Graph-based Text Classification: Learn from your Neighbors, SIGIR 2006.

[3] H. Bast, D. Majumdar, R. Schenkel, M. Theobald, G. Weikum: IO-Top-k: Index-access Optimized Top-k Query Processing, VLDB 2006.

[4] M. Bender, S. Michel, P. Triantafillou, G. Weikum, C. Zimmer: Improving Collection Selection with Overlap Awareness, SIGIR 2005.

[5] M. Bender, S. Michel, P. Triantafillou, G. Weikum, C. Zimmer: Minerva: Collaborative P2P Search, Demo Paper, VLDB 2005.

[6] M. Bender, S. Michel, P. Triantafillou, G. Weikum: IQN Routing: Integrating Quality and Novelty in P2P Querying and Ranking, EDBT 2006.

[7] M. Bender, S. Michel, P. Triantafillou, G. Weikum, C. Zimmer: P2P Content Search: Give the Web Back to the People, IPTPS 2006.

[8] M. Bender, S. Michel, P. Triantafillou, G. Weikum: Global Document Frequency Estimation in Peer-to-Peer Web Search, WebDB 2006.

[9] K. Berberich, M. Vazirgiannis, G. Weikum: Time-aware Authority Ranking, Internet Mathematics 2(3), 2006.

[10] K. Berberich, S. Bedathur, M. Vazirgiannis, G. Weikum: BuzzRank . . . and the Trend is your Friend, WWW 2006.

[11] W.W. Cohen: MinorThird, http://minorthird.sourceforge.net/

[12] H. Cunningham: ANNIE – a Robust Cross-Domain Information Extraction System, http://gate.ac.uk/ie/annie.html

[13] R. Fagin, A. Lotem, M. Naor: Optimal Aggregation Algorithms for Middleware, Journal of Computer and System Sciences 66(4): 614-656, 2003.

[14] S. Golder, B. Huberman: The Structure of Collaborative Tagging Systems, Journal of Information Science, 2006.

[15] J. Graupmann, R. Schenkel, G. Weikum: The SphereSearch Engine for Unified Ranked Retrieval of Heterogeneous XML and Web Documents, VLDB 2005.

[16] G. Ifrim, G. Weikum: Transductive Learning for Text Classification using Explicit Knowledge Models, PKDD 2006.

[17] INEX 2006: Initiative for the Evaluation of XML Retrieval, http://inex.is.informatik.uni-duisburg.de/2006/

[18] J. Lafferty, D. Sleator, D. Temperley: Link Grammar, http://www.link.cs.cmu.edu/link/

[19] J. Li, B.T. Loo, J.M. Hellerstein, F. Kaashoek, D.R. Karger, R. Morris: On the Feasibility of Peer-to-Peer Web Indexing and Search, IPTPS 2003.

[20] J. Luxenburger, G. Weikum: Exploiting Community Behavior for Enhanced Link Analysis and Web Search, WebDB 2006.

[21] W. Meng, C.T. Yu, K.-L. Liu: Building Efficient and Effective Metasearch Engines, ACM Computing Surveys 34(1): 48-89, 2002.

[22] S. Michel, P. Triantafillou, G. Weikum: KLEE: A Framework for Distributed Top-k Query Algorithms, VLDB 2005.

[23] J.X. Parreira, D. Donato, S. Michel, G. Weikum: Efficient and Decentralized PageRank Approximation in a Peer-to-Peer Web Search Network, VLDB 2006.

[24] R. Schenkel, A. Theobald, G. Weikum: Efficient Creation and Incremental Maintenance of the HOPI Index for Complex XML Document Collections, ICDE 2005.

[25] R. Schenkel, M. Theobald: Structural Feedback for Keyword-Based XML Retrieval, ECIR 2006.

[26] S. Siersdorfer, S. Sizov: Restrictive Clustering and Metaclustering for Self-organizing Document Collections, SIGIR 2004.

[27] S. Siersdorfer, S. Sizov, G. Weikum: Goal-oriented Methods and Meta Methods for Document Classification and their Parameter Tuning, CIKM 2004.

[28] S. Siersdorfer, S. Sizov: Automatic Document Organization in a P2P Environment, ECIR 2006.

[29] S. Sizov, M. Theobald, S. Siersdorfer, G. Weikum, J. Graupmann, M. Biwer, P. Zimmer: The BINGO! System for Information Portal Generation and Expert Web Search, CIDR 2003.

[30] R. Steinmetz, K. Wehrle (Editors): Peer-to-Peer Systems and Applications, Springer, 2005.

[31] F. Suchanek, G. Ifrim, G. Weikum: Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents, KDD 2006.

[32] A. Theobald, G. Weikum: Adding Relevance to XML, WebDB 2000.

[33] A. Theobald, G. Weikum: The XXL Search Engine: Ranked Retrieval of XML Data using Indexes and Ontologies, EDBT 2002.

[34] M. Theobald, G. Weikum, R. Schenkel: Top-k Query Evaluation with Probabilistic Guarantees, VLDB 2004.

[35] M. Theobald, R. Schenkel, G. Weikum: An Efficient and Versatile Query Engine for TopX Search, VLDB 2005.

[36] M. Theobald, R. Schenkel, G. Weikum: Efficient and Self-Tuning Incremental Query Expansion for Top-k Query Processing, SIGIR 2005.

A Report on the First International Workshop on Best Practices of UML (BP-UML'05)

Juan Trujillo
Dept. of Language and Information Systems
University of Alicante, Spain
Apto. Correos 99. E-03080
jtrujillo@dlsi.ua.es

1. Introduction

The Unified Modeling Language (UML) has been widely accepted as the standard object-oriented (OO) modeling language for modeling various aspects of software and information systems. The UML is an extensible language, in the sense that it provides mechanisms to introduce new elements for specific domains if necessary, such as web applications, database applications, business modeling, software development processes, data warehouses and so on. Furthermore, the latest work of the Object Management Group (OMG) on UML [1] resulted in a larger and more complicated specification, with even more diagrams for some good reasons. Although providing different diagrams for modeling specific parts of a software system, not all of them need to be applied in most cases. Therefore, heuristics, design guidelines, and lessons learned from experiences are extremely important for the effective use of UML and to avoid unnecessary complication.

This report focuses on the First International Workshop on Best Practices of UML (BP-UML'05) held in conjunction with the 24th International Conference on Conceptual Modeling (ER'05) in Klagenfurt, Austria, on October 24th-28th, 2005. A summary of the accepted papers is given.

In the call for papers, papers focused on the application of the UML in new domains were especially encouraged. In response to the call for papers, the workshop received 25 submissions and only 9 papers were selected by the Program Committee, making an acceptance rate of 36%.

The accepted papers were organized in three different sessions: (i) Experience reports and new applications, (ii) Model evaluation and requirement's modeling, and (iii) Metamodeling and Model Driven Development. In the first one, two papers present valuable experience reports and another one describes how to apply UML for multidimedia modeling. In the second one, one paper is focused on evaluating the cardinality interpretation by users in a UML class diagram, and the other two papers are focused on the Use case diagrams of the UML. Finally, in the third session, while one paper presents how to analyze the consistency of a UML diagram, the other two are focused on the Model Driven Architecture (MDA) and metamodeling. The workshop proceedings are published in [11].

2. Experience reports and new applications

B. Dobing and J. Parsons [2] argue that although many research papers and text books have been published on the different aspects of UML, works on the practical use of UML are absolutely missing. In this paper, authors report results of a survey of UML use by practitioners. Authors developed the survey through the Web and based on a literature review and preliminary interviews with about a dozen practitioners. The Object Management Group (OMG) supported this survey and most practitioners were associated with the OMG. Results indicate varying levels of use, and perceived usefulness, of different UML diagrams such as Use Case, Activity, Sequence, Class, Collaboration and Statechart Diagrams. The reported survey received 299 usable responses, which either contained data on UML component usage (182) or reasons why the UML was not being used (117). Of the 182 analysts using UML components, most (171) were using the UML while 11 were using several UML components as part of another methodology. Another conclusive aspect that the authors argue is that UML is also used by non-IT professional, and therefore, UML diagrams should be more readable and easy to understand.

J.A. Cruz et al. [3] present an interesting approach on the understandability of UML statechart diagrams, and in particular, on how the use of Composite states affects the understandability of these diagrams. To this aim, authors define a new metric named the Nesting Level in Composite States (NLCS) which indicates the maximum number of nested composite states in a UML statechart diagram. Then, the authors focus on describing the experimental process accomplished in order to check the empirical validation of the proposed metric. Unfortunately, the obtained results were not highly conclusive and the authors have not been able to find an optimal use of nesting within UML statechart diagrams and they can only partially conclude that a flat nesting level (0 or 1) within a relatively simple UML statechart diagram makes it more

understandable. Obviously, further empirical research is needed, considering more complex UML statechart diagrams.

T. Ignatova and I. Bruder [4] propose a UML framework to derive applications-specific multimedia database models. The authors mainly focus on describing their framework, which allows us to define the core elements of a multimedia database model, such as mediatype - and application- independent structure, content, relationships and operations. Then, the authors also discuss the advantages of using UML for representing multimedia data as well as shortcomings of this approach that should be covered in the future. Also, they describe the utilization of their UML framework for the instantiation of a model for an image database of scanned handwritten music scores. Finally, the authors showed the advantages of the framework, such as the facilitated design and maintenance of the application, and the seamless integration with other applications.

3. Model evaluation and requirement's modeling

G. Poels et al. [5] present an empirical study on the many-to-many relationships with attributes in Class diagrams. Firstly, the authors provide related work and discussions on the pros and cons of objectifying *many-to-many* relationships in Class diagrams. Then, the authors describe an experiment in order to check if the representation chosen for a relationship with attributes affects the ability of model users to understand the information conveyed by a UML class diagram. The authors employed two pairs of class diagrams representing two structural models including one *many-to-many* relationship with attributes. The results presented in the paper indicate that, controlling for cardinality knowledge, business users can better interpret the information that a UML class diagram conveys about a *many-to-many* relationship with attributes if this relationship is represented as an association class instead of an object class. Finally, the authors argue that the implication for establishing 'best practices' in UML modeling is that modelers should refrain from objectifying such relationships if the goal is an effective communication of domain semantics to business users, who are not UML or modeling experts.

J. Goldman and I-Y. Song's paper [6] argues for the need of a structure or framework for organizing a large number of use cases that may be required for modeling complex systems. Thus, the authors start by analyzing five existing use case classification schemas from existing literature. Then, they propose a new additional classification schema based on the system functionalities for classifying and organizing use cases. The authors also propose a straightforward methodology, resting on sequentially answering some simple questions, to determine use case categories to aid analyzers in real-world projects. In order to illustrate the proposed method, the authors present an exercise conducted in a classroom with 31 graduate students in an introductory UML course. Finally, the authors discuss how we can effectively understand categorized use cases in terms of project priority and personnel skills to achieve the best possible allocation of project resources to use case-driven development efforts.

M. Hilsbos et al. [6] present a comparative analysis of the use case relationships discussed in eleven literatures, including the UML 2.0 specification. First of all, and due to the different terms used in the referred literature, the authors provide a common terminology in order to correctly compare and analyze the related work. Then, the authors provide an extensive literature review and present the agreed usages and different proposed view points of the use case relationships, and argue for a logical resolution for each proposal. As a coherent approach for applying use case relationships, the authors proposed three rules derived from the review of the literature and their own experience and illustrate the rules with examples. Their rules are based on the analysis of preconditions, postconditions of use cases, and characteristics of the behaviors being separated. Finally, from the provided analysis, the authors mainly conclude that practitioners should be aware of the nuances of appropriate application of each use case relationship, apply the relationships sparingly, and, when in doubt, develop several alternative models for complex problems.

4. Metamodeling and Model Driven Development

S. Meliá and J. Gomez [7] propose a generic approach called WebSA (Web System Architecture), based on the Model Driven Architecture (MDA) paradigm to design Web applications. Authors start by providing an overview of the WebSA development process and the modeling notation. Their approach is made up of a set of UML architectural models and QVT (Query/View/Transformations) transformations as mechanisms to integrate the functional aspects of the current methodologies with the architectural aspects. In order to illustrate their proposal, the authors use their WebSA approach to tackle the design of the well known J2EE Petstore specification, showing how to integrate functional and architectural aspects in the design of Web applications. Then, the authors explain the QVT transformations showing how traditional Web functional models and the Configuration model can be merged into an Integration model. Finally, the authors provide an overview of this Integration model.

F.J. Lucas and A. Toval [8] present a rigorous approach to improve the consistency analysis between UML diagrams. To start with, the authors provide a summary of the algebraic formalization of part of the UML metamodel on which their rigorous approach is based. The authors

argue that their framework helps to guarantee the consistency of models because all the specifications are integrated within the same formalism. Then, the authors show the applicability of their approach by verifying the consistency between Class Diagrams and Communication Diagrams. Finally, the authors focus on verifying two properties: (i) a syntactic verification through associations, and (ii) a type consistency of the parameters in the calls of methods.

B. List and B. Korherr [9] propose a UML 2 Profile for Business Process Modelling (BPM). The authors start by discussing the main requirements that a Business Process model should capture. Then, they describe the meta-model as the basis for the proposed UML 2 profile. This meta-model allows designers to consider two complementary perspectives: (i) the business perspective and (ii) the sequence perspective. The sequence perspective refines the business perspective and describes the detailed flow of the process. The business perspective presents the business process from a wide angle by integrating aspects like goals, customers, deliverables, process types etc. Then, the authors provide the specification of the proposed UML 2 profile by defining the required stereotypes, tagged values and constraints. Finally, in order to demonstrate the practical applicability of the business perspective of the UML 2 profile for BPM, the authors apply their profile to *Processing of Claims* business process of an insurance company.

5. Conclusions / Summary

BP-UML'05 was organized on the basis that although UML provides different diagrams for modeling specific parts of a software system, not all of them need to be applied in most cases. Furthermore, due to the considerable number of different diagrams that can be used for modeling the different aspects of a software system, many inconsistencies may appear between the different used UML diagrams. In this workshop, some experimental works were presented in order to help us understand where and when to use the different UML diagrams. Other works showed us how to correctly use the Use Case diagrams and to avoid the inconsistency between different UML diagrams. Finally, another group of papers showed how to apply and extend UML to new applications such as Multimedia, Web or Business Process modeling. In other words, the workshop was a valuable forum where UML researchers will find interesting papers in order to improve the way we can apply UML to real world projects.

Thanks to the number of submissions of this first edition (25) together with the high quality of the accepted papers and the low acceptance rate (36%), it is my pleasure to announce that the second edition of BP-UML is held together with ER2006. My intention is to keep organizing this workshop several years as there is a wide agreement in the UML research community that we still need more and best practices of UML in order to correctly use and apply UML.

6. Acknowledgments

I would like to express my gratitude to the Program Committee (PC) members and the additional external referees for their hard work in reviewing papers. In order to keep the high quality of former workshops held in conjunction with ER, a strong International PC was organized with extensive experience in the UML and their relevant scientific production in the area. I also thank all the authors for submitting their papers and the ER2005 organizing committee for all their support. Another gratitude is for our technician Miguel A. Varo, who developed the web site and the review system (http://gplsi.dlsi.ua.es/congresos/bpuml05/). This workshop was organized within the framework of the following projects: MESSENGER (PCC-03-003-2), METASIGN (TIN2004-00779), DADASMECA (GV05/ 220) and DADS (PBC-05-012-2).

7. References

[1] OMG. UML 2.0 Superstructure and Infrastructure Specifications. http://www.uml.org/#UML2.0. 2005

[2] B. Dobing, J. Parsons. "Current Practics in the Use of UML", in [11].

[3] J. A. Cruz-Lemus et al. "An Empirical Study of the Nesting Level of Composite States within UML Statechart Diagrams", in [11].

[4] T. Ignatova. "Utilizing a Multimedia UML Framework for an Image Database Application", in [11].

[5] G. Poels et al. "Object Class or Association Class? Testing the User Effect on Cardinality Interpretation", in [11].

[6] J. L. Goldman, I-Y. Song. "Organizing and Managing Use Cases", in [11].

[7] M. Hilsbos et al. "A Comparative Analysis of Use Case Relationships", in [11].

[8] S. Melia, J. Gomez. "Applying transformations to Model Driven Development of Web applications", in [11].

[9] F. Lucas, A. Toval. "A Precise Approach for the Analysis of the UML Models Consistency", in [11].

[10] B. List, B. Korherr. "A UML 2 Profile for Business Process Modelling", in [11].

[11] J. Akoka, et al. (eds.). "ER Workshops 2005", Lectures Notes in Computer Science, vol. 3770, pp. 1-96, 2005, Springer-Verlag, Berlin Heidelberg 2005.

Report on the International Provenance and Annotation Workshop (IPAW'06) 3-5 May 2006, Chicago

Rajendra Bose
University of Edinburgh

Ian Foster
Computation Institute
University of Chicago and
Argonne National Laboratory

Luc Moreau
University of Southampton

1. BACKGROUND

The *provenance* of a data item refers to its origins and processing history, while *annotation* is a term that refers to the process of adding notes or data to an existing structure. Because these terms are broad, and are used in slightly different ways by different communities, confusion is rampant. For example, consider that (1) annotating a data set with its provenance information, and (2) finding the provenance of a specific data annotation are both perfectly reasonable concepts.

To help clarify these issues and advance techniques to capture data provenance and facilitate annotation, the International Provenance and Annotation Workshop (IPAW'06) was held May 3-5, 2006 at the University of Chicago's Gleacher Center in downtown Chicago; it was co-chaired by Luc Moreau (University of Southampton) and Ian Foster (University of Chicago and Argonne National Laboratory) and included roughly 45 participants, representing about 25 organizations or projects. The workshop provided some continuity to two earlier events, the Workshop on Data Derivation and Provenance organized by Peter Buneman and Ian Foster in Chicago in 2002, and the Workshop on Data Provenance and Annotation organized by Dave Berry, Peter Buneman, Michael Wilde, and Yannis Ioannidis in Edinburgh in 2003; see Section 4 for brief notes on these previous meetings.

The single-track set of sessions during IPAW'06 [1] consisted primarily of presentations of a selection of papers refereed by the program committee, which will be published as Lecture Notes for Computer Science (LNCS) Volume 4145. The program also included two keynote talks, a discussion regarding the pros and cons of standardizing approaches for capturing and managing data provenance, an entertaining "Gong Show" to foster new and original ideas, and a wrap-up discussion about future meetings and collaborative efforts of this new and growing research community.

2. KEYNOTES AND DISCUSSIONS

In the first keynote presentation, Roger Barga discussed research at Microsoft aimed at supporting scientific workflow creation, featuring the automatic capture of provenance information and the ability to retrieve this information at different levels of granularity. In the second keynote, Juliana Freire (University of Utah) described the Vistrails system for creating versioned visual pipelines to construct scientific visualizations. Derived pipeline versions are connected by "trails"—trees that can be queried and displayed graphically.

Jim Myers (NCSA, University of Illinois, Urbana-Champaign) and Luc Moreau debated and sought audience input on whether the time was right to discuss standard models of data provenance or standard interfaces for recording, querying, and administering provenance stores. The wrap-up discussion at the end of the workshop took up this thread again, with general agreement voiced by the audience to begin a mailing list and continue the growth of this community by keeping the workshop an annual event. Participants also agreed on tentative steps to set up a "Provenance Challenge," which would include a data provenance-tracking scenario and some evaluation measures; these will enable different groups to test and compare their approaches for this scenario during the next several months (See [2]).

A mid-workshop diversion was provided by the Gong Show of outlandish, "outside-the-box" ideas, chaired by Ian Foster. Highlights included presentations tentatively linking data provenance to shoe shopping, horoscope consultation based on the time data was created, the social communication traits of 14-year old girls, divining research funds, selecting breakfast cereal, eschewing junk mail, and "date provenance."

3. SUMMARY OF SESSIONS

The three-day workshop included presentation sessions on applications and systems, semantics, workflow, and models of provenance, annotations and processes. The following sections present brief encapsulations of the presentation topics for each day, and are intended to provide a short overview of the workshop and direct interested readers to the full papers.

3.1 Day One Presentations

In the *Applications* session, Dimitri Bourilkov (University of Florida) spoke about a project to facilitate automatic logging and reuse of data analysis sessions at the UNIX command line by combining the functionality of data analysis software for high energy physics, CVS, and his CODESH UNIX shell with virtual logbook capabilities. Javier Vazquez-Salceda (Universitat Politecnica De Catalunya) discussed applying the Provenance Aware Service Oriented Architecture (PASOA) and EU Provenance projects to the domain area of distributed medical applications, using the example of organ transplant management. Guy Kloss (German Aerospace Center (DLR)) explained how to implement "provenance-awareness" for aerospace engineering simulations. Nithya Vijayakumar (Indiana University) spoke about provenance tracking for near-real time stream filtering within the Calder data stream processing system. Miguel Branco (CERN/University of Southampton) discussed implementing the PASOA model on the Grid for the high energy physics experiment results delivered by the future ATLAS detector at CERN.

3.2 Day Two Presentations

During the *Semantics1* session, Joe Futrelle (NCSA, University of Illinois, Urbana-Champaign) described a system that harvests provenance in the form of RDF triples augmented with actor and timestamp information. Tara Talbott (Pacific Northwest National Laboratory) described how a parser for extracting XML scientific data format descriptions, as well as the data itself, assists with recording provenance. Jennifer Golbeck (University of Maryland) first presented colleagues' work on a web portal for managing images that can be annotated with semantic descriptions; these semantic annotations can help track the provenance of images. She also discussed her project on exploring how annotations in social networks on the web can help record and infer levels of trust. Ewa Deelman (University of Southern California) presented work on augmenting existing metadata catalogs with semantic representations; she described a prototype that allows queries on temporal attributes expressed in OWL.

In the *Workflow* session, Ian Wootten (Cardiff University) spoke about using a prototype to explore how to capture actor state assertions during the enactment of a process according to PASOA ideas. Ilkay Altintas (San Diego Supercomputing Center/University of California, San Diego) discussed a provenance framework for the open source-based Kepler scientific workflow system; this framework includes a provenance "listener" utility to save information about the details of workflow executions. Bertram Ludascher (University of California, Davis) noted the importance of accounting for different models of computation, such as the directed acyclic graph (DAG), process network (PN), and synchronous data-flow (SDF) models, in constructing user-oriented provenance for pipelined scientific workflows. Michael Wilde (University of Chicago/Argonne National Laboratory) discussed provenance collection for large-scale workflow execution in the Virtual Data System (VDS), and work on providing the ability to query virtual data relationships, annotation and workflow patterns.

Finally, in the *Models of Provenance, Annotations and Processes* session, James Cheney (University of Edinburgh) presented a formal model for the process of manually curating databases with cut and paste operations. Margo Seltzer (Harvard University) described the idea of storage systems that automatically collect complete, low-level, and query-able data transformation details. Simon Miles (University of Southampton) discussed the idea of using a filter to provide the proper scope for provenance queries on potentially large directed acyclic graphs. Rajendra Bose (University of Edinburgh) presented work on a prototype system that allows individual research groups to create annotations over existing, distributed catalogues of astronomy data; these annotations record a group's assertions of matching entries across the different catalogues.

3.3 Day Three Presentations

The *Systems* session began with Victor Tan (University of Southampton) speaking about security issues within the PASOA model; he discussed approaches to achieving access control on potentially sensitive combinations of assertions within a provenance store. Jane Hunter and Imran Khan (University of Queensland) described a system architecture that combines existing Semantic Web annotation (Annotea) and security (Shibboleth, XACML) components. Yogesh Simmhan (Indiana University) presented a quantitative performance comparison of two methods of recording provenance for scientific workflow execution: the Karma

framework and the Provenance Recording Protocol (PReP) from the PASOA project. Christine Reilly (University of Wisconsin) discussed how to achieve provenance functionality for a distributed job execution system like Condor. Last, Ludek Matyska (CESNET) discussed how to enhance basic Grid job tracking with more detailed job provenance in the gLite middleware developed as part of the EU Enabling Grids in E-SciencE (EGEE) project.

During the *Semantics2* session, Carole Goble (University of Manchester) discussed the "identity crisis" caused by the assignment of multiple identifiers to the same entities within bioinformatics workflows; she proposed using sets of IDs to manage this potential problem for determining an entity's provenance. Hugo Mills (University of Southampton) spoke on behalf of David De Roure about the Combechem project which uses an RDF store to capture the provenance and semantics behind human-driven experimental chemistry, including the precise environmental conditions during an experiment. In the final talk, Paul Groth (University of Southampton) expounded on two principles for high quality documentation of provenance, which he explained were supported by PReP: only recording facts that (1) can be verified and (2) possess correct attribution.

4. PREVIOUS WORKSHOPS

We close this event report with some brief notes on previous related workshops that helped to set the scene for IPAW'06.

The Workshop on Data Derivation and Provenance organized by Peter Buneman and Ian Foster in Chicago in 2002 [3] was an important first venue for comparing and contrasting the definitions, expectations and requirements of data provenance and annotation from those involved with data management across various scientific domains. Most participants submitted position papers, and a number of short presentations were given [4]. The Workshop on Data Provenance and Annotation organized by Dave Berry, Peter Buneman, Michael Wilde, and Yannis Ioannidis in Edinburgh in 2003 [5] included surveys of provenance topics selected by the organizing committee with short presentations [6].

The 2002 provenance workshop had about 40 participants representing roughly 15 organizations or projects, and the 2003 provenance workshop had about 50 participants representing roughly 20 organizations or projects. About 10 of the same organizations/projects attended both workshops. The topic of workflow was considered in the 2002 provenance workshop, while the next year a separate e-Science Workflow Workshop [7] with 75 participants immediately followed the 2003 provenance workshop, with some participants attending both meetings.

5. ACKNOWLEDGMENTS

IPAW'06 was sponsored by Springer and Microsoft, and endorsed by the Global Grid Forum. Acknowledgment and thanks are due to the IPAW'06 program committee, and the hosts and meeting coordinators at the University of Chicago and Argonne National Laboratory.

6. WEBSITE REFERENCES

[1] http://www.ipaw.info/ipaw06

[2] http://twiki.ipaw.info/bin/view/Challenge

[3] http://www-fp.mcs.anl.gov/~foster/provenance

[4] http://people.cs.uchicago.edu/~yongzh/position_papers.html

[5] http://www.nesc.ac.uk/esi/events/304/

[6] http://www.nesc.ac.uk/action/esi/contribution.cfm?Title=304

[7] http://www.nesc.ac.uk/esi/events/303/

Report on SciFlow 2006: The IEEE International Workshop on Workflow and Data Flow for Scientific Applications

Brian F. Cooper
College of Computing
Georgia Institute of Technology
cooperb@cc.gatech.edu

Roger Barga
Microsoft Research
barga@microsoft.com

1. Introduction

Computation has been described as the "third leg" of science, along with theory and experimentation. Certainly, modern information systems are vital to managing and processing the huge amounts of data produced by simulations and experiments. However, existing tools are only now beginning to catch up with the needs of today's scientists, and much more needs to be done to support the computational needs of tomorrow's scientists. In particular, scientists still need effective tools to deal with massive data sets that may be geographically scattered, to apply multiple complex and interacting transformations to the data, and to ensure the quality and repeatability of their computations. The IEEE SciFlow workshop brought together computing researchers who are exploring how to build the next generation of information systems to address these needs. The workshop was held on April 8, 2006, in conjunction with the IEEE International Conference on Data Engineering in Atlanta, Georgia, USA.

The papers presented in this workshop demonstrate the ability of computer scientists and natural scientists to work together to create computer systems that support scientific exploration. The workshop itself was very interactive, with the audience raising many questions for the speakers and different speakers adapting their talks to address points brought up in discussions. This interaction was greatly helped by generous sponsorship from Microsoft, which provided a lunch and cocktail reception so that participants could continue their discussion.

2. Workshop themes

Several overall themes emerged from the discussions. One theme is that although many groups are building systems today, there are still many open research problems and a lack of standard tools for use by scientists. There are several available tools for workflows (such as the Kepler workflow system, Windows Workflow Foundation and components of IBM's WebSphere), but work still needs to be done to adapt many tools to the scientific domain, and make them usable by non-computer scientists. In addition to tool development, important research problems include:

- Applying "general purpose" tools to problems with very specific requirements and unique needs
- Adapting workflow and dataflow techniques for vastly different scales (from individual laboratory information systems all the way up to large multi-national collaborations)
- Managing the quality and provenance of information; for the scientific data itself, for the workflow specifications (and their various versions), and for data products and visualizations of the data

Another theme was that these systems tend to be built in isolation with little learning or re-use from other projects. Although many systems are currently being built, there is not an easy way for one group to learn about what is being done by other groups, unless there happens to be a person in common with both projects. Also, it is difficult for a group that is beginning to develop a workflow for a particular application to learn about the universe of tools and solutions that are available. It might be useful to have some repository of "best practices" or "experiences" for developing workflows, so that developers do not have to start from scratch each time.

A third theme was that developers of "new and exciting" solutions need to be willing to work with scientists and use their existing legacy tools and processes. Scientists become very attached to the GUIs they know how to use, to the information systems they have already spent time and resources developing, and to specialized codes (often written in FORTRAN) that they trust. Insisting that an application be ported to a new language or provide a new interface is often not feasible. Thus, much of the challenge in building these systems is to retain the components that scientists want to keep, while connecting them in new ways to facilitate better and more interesting functionality.

3. Papers and presentations

The workshop program included 10 papers, which can be roughly categorized into papers on "tools" and papers on "case studies."

Several case studies of developing requirements and systems for specific applications were presented. These case studies illustrate how widely useful workflow and data flow systems are in modern science. Scott Klasky of Oak Ridge National Laboratory discussed requirements for analyzing plasma simulation data, and focused on the need for flexibility in adapting and deploying workflows of parallel physics codes. Laura Bright of Portland State University discussed applying the "factory" metaphor to managing large numbers of data-product-generation workflows to produce as many data products as possible within an allotted time period. The key challenge is to make the best use of the "factory floor" (e.g., the available high performance computing resources) to produce as many data products as possible. Simon Cox of the University of Southampton described using the Windows Workflow Foundation product to manage wind-tunnel experiments, and discussed analyzing and visualizing data in real-time (so that problems can be detected quickly to avoid wasting an experimental run.)

Mirek Riedewald of Cornell University reported on experiences managing several different large-scale data flows. One such data flow is sky survey data from the Arecibo telescope, where the main challenge is efficiently dealing with large amounts of data. Another data flow results from the CLEO high energy particle physics experiments, which required adding provenance support to a large body of legacy code. The third data flow is a large collection of World Wide Web data (the Web Lab) for sociological studies, where the main challenge is navigating multiple large snapshots of the web.

In terms of tools, several groups are building general purpose or generalizable tools, although in many cases these tools are motivated by specific applications. Louiqa Raschid of the University of Maryland described adapting an enterprise-style mediator system, DB2 and WebSphere, for use in scientific applications. The key idea is to express the workflow as a large SQL query, and then utilize the mediation capabilities to efficiently execute and monitor the workflow. Bettina Kemme of McGill University described Exp-WF, a workflow system for managing data in laboratory-scale information systems. Her experience demonstrated the importance of keeping some legacy components (e.g., the laboratory information systems themselves), and using a paradigm of "add to, don't replace" when developing the system's functionality. Bertram Ludäscher of the University of California at Davis argued that dataflow process networks are a natural model for specifying data-intensive scientific workflows, but that control-flow and plumbing-intensive tasks lead to "messy" dataflow designs. He proposed an approach that allows one to nest state-machines (for flexible control-flow) within dataflow networks, resulting in simpler, more reusable workflows. Reusability is further enhanced by a "workflow template" mechanism. Louiqa Raschid (presenting on behalf of Zoe Lacroix from Arizona State University) described the SemanticBio system, where the workflow can be specified using high-level ontologies, separating the specification from the implementation details.

Workflow tools can provide additional value to scientists by helping them manage the provenance and quality of their data (and of the workflows themselves). Juliana Freire of the University of Utah described how scientists often spend many hours tuning and tweaking visualization workflows to provide just the right view of their data. The VisTrails system, that she and her colleagues are developing at Utah, maintains detailed information about the exploratory process - the trial-and-error steps followed to construct a set of data products. By capturing the provenance of both the derived data and the processes that generate these data, and by providing an intuitive interface for comparing the results of different workflows, VisTrails greatly simplifies the scientific discovery process. For example, it allows a scientist to manage the versions of their visualization workflows, reverting to an earlier version if necessary, comparing different visualizations side by side, and so on. Yogesh Simmhan of Indiana University described a quality model for collaborative data that allowed scientists to evaluate data on multiple axes: the metadata, the provenance, the quality of service when accessing the data, and the community evaluation of the data.

4. Moving forward

Certainly, scientific workflow and data flow systems will continue to be built, as scientists increasingly recognize their usefulness. The challenge for computer scientists is to develop tools and techniques that ease the process of creating, maintaining and executing workflows, and allow scientists to focus their energies on the science, and not on the "plumbing."

As mentioned above, building a repository or forum for sharing experiences and best practices for these systems would be a significant help to the community. One participant noted that just having people who have actually built these systems in the same room to discuss their experience is a great way to find out about new tools and avoid repeating mistakes.

In particular, sharing "tips and tricks" is important for at least two levels: the "IT" level, where computing centers are deploying tools for use by their scientists, and at the "research level," where computer scientists are developing new techniques and algorithms based on real requirements from natural scientists. One possibility for such a forum is to repeat this workshop; another possibility is to form some sort of working group that could sponsor a portal, newsletter or other forum.

Also, the need to integrate heterogeneous data, heterogeneous systems, new and legacy codes, and so on means that any tools and techniques developed in the future need to "play nicely with others" in order to have a realistic chance of adoption. Forcing scientists to use a particular programming language, operating system or data format is infeasible, given both the existing base of software and the specialized hardware requirements of various groups.

Acknowledgements

We would like to thank the authors for their compelling papers and interesting work as well as the lively discussion at the workshop. We would like to thank the program committee for providing high quality reviews, even over their holiday break. We would like to thank the ICDE organizers for providing us with a venue. We would finally like to thank Microsoft for their support of the workshop and for their commitment to help move this area forward.

The SciFlow 2006 webpage is at http://www.cc.gatech.edu/~cooperb/sciflow06/.

Papers

[1] William Y. Arms, Selcuk Aya, Manuel Calimlim, Jim Cordes, Julia Deneva, Pavel Dmitriev, Johannes Gehrke, Lawrence Gibbons, Christopher D. Jones, Valentin Kuznetsov, Dave Lifka, Mirek Riedewald, Dan Riley, Anders Ryd, Gregory J. Sharp. "Three Case Studies of Large-Scale Data Flows."

[2] Shawn Bowers, Bertram Ludaescher, Anne H.H. Ngu, Terence Critchlow. "Enabling Scientific Workflow Reuse through Structured Composition of Dataflow and Control-Flow."

[3] Laura Bright, David Maier, Bill Howe. "Managing the Forecast Factory."

[4] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Claudio T. Silva, Huy T. Vo. "Managing the Evolution of Dataflows with VisTrails."

[5] Barbara A. Eckman, Terry Gaasterland, Zoe Lacroix, Louiqa Raschid, Ben Snyder, Maria Esther Vidal. "Implementing a Bioinformatics Pipeline (BIP) on a Mediator Platform: Comparing Cost and Quality of Alternate Choices."

[6] Brian Gabor, Bettina Kemme. Exp-WF: Workflow Support for Laboratory Information Systems.

[7] Scott A. Klasky, Bertram Ludaescher, Manish Parashar. "The Center For Plasma Edge Simulation Workflow Requirements."

[8] Herve Menager and Zoe Lacroix. "A Workflow Engine for the Execution of Scientific Protocols."

[9] A. Paventhan, Kenji Takeda, Simon J. Cox, Denis A. Nicole. "Leveraging Windows Workflow Foundation for Scientific Workflows in Wind Tunnel Applications."

[10] Yogesh L. Simmhan, Beth Plale, Dennis Gannon. "Towards a Quality Model for Effective Data Selection in Collaboratories."

Jennifer Widom Speaks Out

on Luck, What Constitutes Success, When to Get Out of an Area, the Importance of Choosing the Right Husband, Outlandish Vacations, How Hard It Is to Be an Assistant Professor, and More

by Marianne Winslett

Jennifer Widom
http://infolab.stanford.edu/~widom/

Welcome to this installment of ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today we are at the Department of Computer Science at the University of Illinois at Urbana-Champaign. I have here with me Jennifer Widom, who is a professor of Computer Science at Stanford University. Jennifer's research currently focuses on data provenance, management of uncertainty, queries over web services, and data streams. Jennifer is a member of the National Academy of Engineering, is an ACM Fellow, and is a former Guggenheim Fellow. Before joining Stanford, Jennifer worked at IBM Almaden Research Center. Jennifer's PhD is from Cornell. So, Jennifer, welcome!

Thank you!

Jennifer, you changed the name of your group from "The Database Group" to "The InfoLab". Is that because database research is dead?

No, it's not, of course. I think of myself, actually, as doing core database research. Most of my students do core database research, and will for a long time to come. But we do have students in our group who work outside the database area. The name change was primarily so students didn't get pigeonholed inappropriately. For example, we had a student who did a thesis in photo browsing and labeling. We didn't want that student to be labeled as a database student. We have students who do information retrieval; again, we don't want them to be pigeonholed as database researchers. Database research is a subset of the work going on in the InfoLab.

You seem to have a knack for picking up-and-coming research areas and being among those who lead the charge. What is your process for choosing problems to work on?

More or less random! When we started working on the LORE project (a database management system for semi-structured data), it was a small offshoot of a data integration project that was using a semi-structured data model. I said, why don't we build a system to store and query that data? And we did, and it grew into a big project, eventually becoming a database system for XML.

I also had a project on data streams. I had wanted to work on data streams for a long time; I thought it was an interesting new model, but I couldn't convince any students to work on it. Finally I got a couple of students interested and we launched a project. Since then, the data stream area has become very popular.

As for my most recent project, I had been thinking it was time to have a new research direction, but I really didn't have any ideas on which way to go. I was doing my morning jog one day when I started to think about uncertainty and data lineage, or provenance, and how they seem to work together in a lot of applications. I decided to build a system to handle these two aspects of data, together with the data itself.

So I would say all of these research directions are chosen fairly randomly, almost as knee-jerk decisions, without a major thought process, or a huge vision. I don't think of myself in fact as a visionary whatsoever.

How do you decide when it is time to leave your current area of research and move into a new one?

I probably do leave areas fairly early, in a sense. I definitely leave and move on to something new when there is still lots of work to do in the old area. There are a couple of driving factors in play here. The first is graduate student interest. Suppose that I have been working on a project for several years and I have a fairly mature prototype. Then even if I have a list of five obvious thesis topics, if I get a first year PhD student, they won't want to work on one of those topics. The area is now four or five years old, and the students don't want to work on a well-defined topic in an "old area". So the students themselves drive approximately a five year cycle in my projects.

The second factor is that if an area gets to have lots of people working in it, I prefer to move on to a newer area with fewer people working in it. I like to do things early and then move out. Sometimes I think of an analogy of surfing: you're riding a wave and then at some point you just cut out and let the wave continue. I like to do that.

Do you have any tips for us on starting a research project in a new area?

To start a project in a new area, I recommend that you spend a year or so on foundations and figuring out how the new area fits in the scheme of things. If you are working in a new area, you can spend a long time working out foundations. I think foundations are important, but sometimes you just have to say that it is time to push on and start building a system, looking at more practical issues. Get some applications, some data, and work with those.

What do you view as your most successful research project to date?

I think that a project of mine is successful if people in industry get interested and start building similar things. Right now I feel that the data streams project has been the most successful. Industry has gotten very interested in the area. People are interested in what we did in the project, and in the query language we developed.

The other project I feel most strongly about is the LORE project, where we built a system that was eventually used for XML. That project got very well known because of XML, and I think it was lucky timing. We had a database system we were building for a semi-structured data model, and along came XML, which was very close to our model. We switched our model to XML, which didn't take much effort at all, and suddenly we had a system for XML very early. So I think the LORE project was also quite successful, and some lucky timing was involved in its success. Lucky timing is not such a factor in the success of the streams project.

So are you saying the industry is taking its cue from what's happening in academia?

No, I don't sense that's happening. I think industry finds business needs, and then they look for the technology that meets those business needs. So that's what I am finding in streams: people from different areas are saying that they need technology that looks like data streams, looks like continuous queries. They discover the technology after they identify what kind of technology they need.

What about what David DeWitt says, that we don't need specialized databases for stream data management?

That may be true, and I'm not going to argue that we absolutely have to have a special purpose data streams system. Our project produced a data stream management system, built from scratch, and that was a lot of fun, but is that going to be the way data streams come to the mainstream, so to speak? Not necessarily.

There are companies building systems that are basically what ours was, a native data stream system. There are companies whose product is based on data streams and continuous queries hidden inside the software. And I have very little doubt that the major DBMS vendors are going to add stream technology to their systems. Other technologies in databases have gone this way, too. At first, little startups pop up, building native systems. XML databases went this way. Object-oriented databases went this way. I am sure there are many other examples. But then the big vendors said, "Hey, we can add that to our systems. It's not that hard, it's seamless with what we already have." Then the little companies get snuffed out. I wouldn't be surprised to see that happen with streams technology. I think the verdict is still out, so I'm not going to argue with the claim that we don't need a special purpose data stream management system.

A different argument might be that we don't need data stream technology at all. I disagree with that. I think that in the last year or two, industry has shown that you do need a different way of looking at streaming data, and you need continuous queries.

So what's the killer application that has emerged from industry?

Financial monitoring is one of the major applications right now. Another is web click data streams, which tell where people are going on the web, and doing things with those streams in real time. There is an application called business activity monitoring, where people want to record everything that is going on in their business in a streaming fashion, and have dashboards where they can view it and allow their high-level executives to make decisions. Those are three real applications that have a real need for this kind of technology.

Where do you think the field is going now?

I can toot my own horn and tell you about my next project, Trio. I'm interested in managing uncertain data, and I do think people are going to want more and more support for uncertainty pushed into the DBMS.

People have worked on uncertain data for a long time, but it hasn't been a primary focus. When I look around now, people are ready to put information about the certainty of their data into a database and start querying it. That really hasn't happened yet, but when you talk to people, they often tell you that they have data that isn't black and white. So that's what I am interested in right now.

This topic seems to connect well with the area of lineage, because if your data isn't certain, you might want to know where it came from and how it was derived, to try to figure out the quality of the data. Quality is very important.

Now I'm done with tooting my own horn. In terms of other areas, data integration will continue forever---people will always work on some problem in data integration. So the database field is always going to have a big component of that. What exactly people will try to do, I'm not sure. I have one student looking at web services and trying to integrate queries across web services.

Data sets are going to get bigger and bigger, messier and messier. Data cleaning is important.

Jennifer, you have spent your whole career in the heart of Silicon Valley, and you've never done a startup. Why is that?

I'm not a startup kind of person. Let me elaborate on that a little. We did come close to doing a startup with the LORE project. But I found that going to meetings with funding people and potential CEOs was not a fun thing for me. It was fun at first, when it was novel, but I lost interest at some point. I also like to be in control of my time and be my own boss. Even if you are the boss of a company, there is always somebody else that is effectively your boss: your funders, or your customers. I wasn't too excited about that. I also like to have predictability, and startups are very unpredictable. For some people, the unpredictability is the joy of doing a startup. For me, that's not a joy. So I've been very content with consulting and sitting on advisory boards. For me, that's the perfect interaction in Silicon Valley. Doing some consulting and talking to people makes me feel that the research I do is grounded, without my having to go for the whole startup package and all that that entails.

You're one of the few people I've interviewed with kids in grade school. How can you be so productive and raise two kids?

I think the most critical factor by far in managing children and having a career is to have the right husband. My husband is also a professor, but we are very, very symmetric about everything, so I have not taken on more burden than him in child raising. So, for those of you watching or reading this, think about that before it is too late!

Having the right husband is the primary factor, but the next most important factor is being efficient, knowing what's important and what's not important, and not being shy about ignoring those things that are not important. Or perhaps it is not *ignoring* so much as *focusing* one's time and energy on what really matters at work, so that you free up time for your family.

How many hours of sleep do you get a night?

I have been known not to get quite enough sleep sometimes. In busy, busy periods I'm a 5-6 hour sleeper. It is definitely important for me to have those extra couple hours during each day for managing the family and my work.

How do you manage to keep your desk so neat?

Those people who have been to my office know that my desk is very neat, and I really love throwing things away. I am a "thrower-awayer". A lot of people are pack-rats, and I am the opposite. I have a philosophy that if you throw everything, or nearly everything, away, the amount of time that you are likely to spend reacquiring something that you threw away and then found that you needed is much, much lower than the amount of time you might spend doing things with those things that you didn't throw away. To put it another way, the time invested in dealing with the case where you accidentally threw something important away is quite low. The time invested in dealing with all the stuff you thought might have been important is quite high. I throw away things in the office, most everything, and even on the computer I throw things away. So my computer desktop is fairly neat.

Last year at SIGMOD, we had a keynote talk called "MyLifeBits, A Transaction Processing Database for Everything Personal," by Gordon Bell from Microsoft (http://research.microsoft.com/barc/mediapresence/MyLifeBits.aspx) . That talk was about the trend to record absolutely everything about one's life, to keep it and have it available. So, for example, let's keep a record of yesterday's phone conversation, just in case we need to look it up someday. I do not want that at all. It's completely against my philosophy to have all that stuff, even if it is supposedly easy to access, because then you spend time trying to access it, and I think that is not a good use of time. I put that all into the efficiency category that I was talking about earlier: trying to live an efficient life so that one has time to work and be with one's family.

Your undergraduate degree is in music. How does a music major end up doing database research?

I was a trumpet performance major at the Indiana University School of Music, where we had the whopping requirement of taking *three* classes that were not music theory or music performance. One of those classes I took was even in the music school, and was called "Computer Applications in Music Research." I chose the course completely randomly as one of my electives. We wrote SNOBOL programs to analyze streams representing music, and I got hooked. It was my junior year in college, and I started taking some computer science classes at Indiana. I finished my performance degree in trumpet, and then I stayed at Indiana. I continued actively in music there, but I switched to a master's program in computer science. That is really how I got my undergraduate education in computer science---I like to think of it that way. They admitted me to computer science at Indiana on the basis of a few classes, and then I really got some breadth, and also got into some research to some extent. After my sort-of-bachelor's degree, which is actually my master's, I decided to get a PhD and moved to Cornell. So that's how I moved from music to computer science. I continued playing my trumpet until 1992, which was quite a while after I finished my PhD.

What made you give up the trumpet?

I was tired of practicing. It's a little bit like a sport, though some people don't realize that. I was practicing my trumpet an hour and a half a day during the time I was working at IBM Almaden, and playing actively in musical groups around San Jose. One day I realized that I didn't really want to do all this practicing any more. I didn't want to just rachet back, because playing the trumpet is such a physical activity. So I just decided to stop. However, I am thinking about taking it up again because my son has just taken up trumpet and he needs someone to play duets with.

You have been both at an industrial lab and at a university. How do you look back on your days at IBM?

The days at IBM were great. They were very easy days. At that time at IBM, we were really just chartered to do research. There weren't a lot of administrative duties; you didn't have to get grants, like a faculty member would have to. We really had a lot of license. I was in a group working on the Starburst project, building a big prototype that was excellent infrastructure for trying out research ideas. I spent most of my time doing research; I believe I really established my research track record during that time at IBM. So my days at IBM were very idyllic.

I also learned about databases mostly at IBM. My PhD is in programming languages, so it was a time to learn about databases, have a lot of freedom, in what was at that time really one of the greatest groups in the world. So it was a great five years I spent there.

What led you to make the switch from industrial research labs to academia?

I am a child of a professor; I always thought being a professor would be a great thing to do. With academia in my genes, the opportunity to have a faculty job at Stanford was something I couldn't turn down.

Would you ever consider moving back to a research lab?

Nope. I love being a professor. It is the greatest job on earth. I believe that to be true.

How does the programming languages community differ from ours?

It's very different. I got my PhD in programming languages and went to some of their conferences for a few years. I find the database community to be friendlier, more social, less self-conscious or posturing. I don't want to put down the programming languages community, but I really find the database community relaxed. I think it may have to do with the funding situation. Now funding is hard to find no matter what community you are in, but there was a period when database people had it quite a bit easier than those in other fields. Also, the database community has a stronger connection to industry, a lot of self-confidence that what they are doing matters to people, and that may be less true in programming languages. There are a lot more party animals in databases, too!

You like to take exotic vacations with your family. (When I was preparing for this interview, some of my informants used words like "outlandish" and "dangerous" to describe your vacations.) How did the vacation turn out where you were going sailing in Thailand with your husband and kids when the tsunami was on the way?

There are a few misconceptions in your question. I'm not going to argue necessarily with "dangerous" and "outlandish"---well, we would not put our family in danger, but perhaps some people think our trips are outlandish. We do take exotic adventure vacations. Another misconception is that---I don't think anyone knew the tsunami was coming, did they? I think it just came. The third misconception is that actually we weren't in Thailand for the tsunami. We were trying to make a reservation to charter a sailboat to sail around some islands in Thailand for that particular time. The sailboat we wanted wasn't available, so we went to New Zealand instead. In fact, we were white water rafting in New Zealand when our guide mentioned to us that a big tsunami had hit in Thailand. We thought, wow, we could have been sailing there! But we weren't, we were just fine. We went back to Thailand the following year, chartered a sailboat and checked out where the tsunami had hit.

Jeff Ullman has a company that makes a system that automates the generation of homework. Do you like using his Gradiance software?

I am a huge fan of Gradiance. (Actually, I argue with Jeff a lot about some aspects of the system, but in reality I love it, so put that in print.) Gradiance generates homework problems from a questions bank, and then it corrects them and gives feedback, all automatically. Students can do their exercises over and over, getting different instances of questions each time, and get it graded immediately. The students love that. Gradiance also has a SQL engine, for introductory database students. You can give Gradiance a schema and data, and assign queries in English. The students write the queries in SQL, and the queries are run against the database, providing immediate feedback. Gradiance will tell the student that they got the wrong answer, and show them the data and the correct query answer. Then the student can try to write the query again. Gradiance has some very clever techniques, like having a second hidden database so students can't fool the system once they see what the query answer is. Students really react positively to the SQL engine

and what it offers. So I have found Gradiance to be an excellent teaching tool, one that I really enjoy using.

I have heard that you think presentation skills are extremely important. Can you expand on this point?

Regarding oral presentation skills, in our group we do have a reputation of having our students give practice talks and then ripping them apart. I think we have very high standards in our group for what it takes to give a good conference talk, and how important it is to give a good talk. So our students do give talks over and over until they get them right. And we do a lot of coaching.

I feel the same way regarding writing skills. I think that writing a clear paper is extremely important and very difficult. We spend quite a bit of time talking about what constitutes a good paper and what doesn't. I am very fussy with my students' drafts. I always warn my younger students that the first time they give me a paper and I mark it up, the student will barely be able see the black ink underneath the blue ink from my pen.

One of my senior students has decided that his papers are *too* well written. He thinks they are so well written that referees have a much easier time finding something to find fault with, because they actually understand the paper. So he is thinking about conducting some experiments where he takes a well written paper, makes it less well written, and submits it to conferences to see if he gets fewer complaints. He really believes this to be the case, and I think there could be some truth to that, unfortunately.

You can get the benefit of the doubt sometimes if the reviewer is not sure what you're talking about.

Exactly. I think that happens quite often. The faster someone reviews a paper, the more it works to your benefit to have not done a good job with your paper.

What's next in your career path? Do you envision yourself as a dean?

At some point way back, I thought it might be fun to be dean. Now I don't think it would be fun, for the same reasons as for a startup: when you are a dean, you start to lose control over things that are important. You lose control over your schedule, you have to dress up all the time (I am not big on dressing up), and you are interacting with people to try to get big donations and such things, which I don't think is my cup of tea. As I said earlier, I am not a visionary, and I think deans ought to be visionaries. So I don't see a deanship as something for me.

Actually, next in my career path, my family is going to take a 14 month trip around the world and not work at all. When I come back from that, I guess I will see how I feel about things.

Do you have any words of advice for fledgling or mid-career database researchers or practitioners?

I can comment on researchers primarily. It is no secret that it is very difficult right now to be a young researcher. I don't think people should pretend otherwise. It is harder now than it used to

be. I think you really have to want it with a passion if you want to be an assistant professor. You have to get grants, which is harder than it used to be. You have to make sure you get major publications, which is harder and more random than it used to be. So it's not easy, and you really do need to want it. My advice is to get that fire going, hit the job with a passion, and don't get discouraged.

I guess "don't get discouraged" is the best advice. You have to look at the bigger picture. If your papers don't get into one conference, it's probably not because you are a terrible researcher. Just wait for the next conference and try to look at the long term, try to make your work have overall impact rather than worrying about each specific instance.

It sounds like you think the acceptance rates at major conferences should be higher.

It's possible. We could go into a whole discussion about conferences and publishing and what's wrong. I think there is a problem, and there has been a lot of discussion over the last couple of years. I'm not sure what the best solution is. Some people have talked about online journals with no acceptance rate, just a pure quality threshold, which I think is an interesting idea. It is a very complicated issue, but I do think that right now in our very selective conferences, many valuable papers aren't being accepted. I do worry about young people's careers because of that. Students also have high pressure now to get papers published, if they are going to look for academic jobs. So it's tough right now.

If you magically had enough time to do one additional thing at work that you are not doing now, what would it be?

I don't have enough time to learn about other areas of computer science, or areas outside computer science. (I suspect that this is the most common answer to this question.) I would love to know much more about all kinds of things. Even closely related things, like information retrieval, data mining---I mean, these are practically in my field and I don't know enough about them. AI, natural language understanding---these are all things I *should* know more about. And then there are things I just *want* to know about. I would like to know more about biology because it's so popular and interesting, graphics, just all kinds of areas.

Having time to learn about those things would be great. I don't foresee it happening. Maybe when my kids go to college.

If you could change one thing about yourself as a computer scientist, what would it be?

I would probably like to do more coding. I think of myself as a systems person, primarily, and my students all build systems. I would like to move down more in my level of interaction with those systems, because I am really at a high level. I would rather know more exactly what is going on inside the system, and even participate myself in building it. That would be great, but I don't have time for it now.

Thank you, Jennifer, for talking with me today.

CALL FOR PAPERS

26th ACM SIGMOD–SIGACT–SIGART Symposium on

PRINCIPLES OF DATABASE SYSTEMS (PODS 2007)

June 11–13, 2007, Beijing, China

`http://sigmod07.riit.tsinghua.edu.cn/` or `http://www.sigmod.org/`

The PODS symposium series, held in conjunction with the SIGMOD conference series, provides a premier annual forum for the communication of new advances in the theoretical foundation of database systems. For the 26th edition, original research papers providing new insights in the specification, design, or implementation of data management tools are called for. Topics that fit the interests of the symposium include the following (as they pertain to databases):

> *algorithms; complexity; computational model theory; concurrency; constraints; data integration; data mining; data modeling; data on the Web; data streams; data warehouses; distributed databases; information retrieval; knowledge bases; logic; multimedia; physical design; privacy; quantitative approaches; query languages; query optimization; real-time data; recovery; scientific data; security; semantic Web; semi-structured data; spatial data; temporal data; transactions; updates; views; Web services; workflows; XML.*

Papers should be at most ten pages, using reasonable page layout and font size of at least 10pt. Additional details may be included in an appendix, which, however, will be read at the discretion of the program committee. *Papers longer than ten pages or in font size smaller than 10pt risk rejection without consideration of their merits.*

The submission process will be through the Web; a link to the conference website will appear on the SIGMOD website (www.sigmod.org) in due time.

Important Dates:

Short abstracts due:	28 November 2006
Paper submission:	5 December 2006
Notification:	26 February 2007
Camera-ready copy:	20 March 2007.

The results must be unpublished and not submitted for publication elsewhere, including the proceedings of other symposia or workshops. All authors of accepted papers will be expected to sign copyright release forms. One author of each accepted paper will be expected to present it at the conference.

Best Paper Award: An award will be given to the best submission, as judged by the program committee.

Best Newcomer Award: There will also be an award for the best submission, as judged by the program committee, written solely by authors who have never published in earlier PODS proceedings.

The program committee reserves the right to give both awards to the same paper, not to give an award, or to split an award among several papers. Papers authored or co-authored by PC members are not eligible for an award.