



データ工学研究専門委員会ニュースレター 創刊

データ工学研究専門委員会 専門委員長からのご挨拶

データ工学研究専門委員会 専門委員長 川越恭二
立命館大学 情報理工学部

2008年度より国内データベース組織が一体となってDB関連の行事を企画し実行していくことになりました。同時に、本研究専門委員会(DE研)もこれまでの研究会中心の活動からより広くDBの方々にとって親しみやすい組織となるようにリニューアルを開始いたしました。今回のニュースレターはその大きな一歩になるものと期待します。第一号の内容はDE研が変化したことを示すようなコンテンツになったと感じます。今後も継続性を第一に考えDE研の現状を踏まえて定期的にニュースレターを発行していきたいと考えております。皆様からのご意見を頂戴できれば幸いです。最後に、第一号の鬼塚編集委員長にはゼロからの立ち上げでご苦労がかなりあったと推察します。ここに厚くお礼申し上げます。

Index

ニュースレター創刊にあたり	2
喜連川先生インタビュー	4
斉藤先生インタビュー：本当に「使える」XMLを目指して	16
SIGMOD2008のProgram Committeeの経験を通して 小杉 尚子	23
データ工学研究専門委員会からのお知らせ	27



ニュースレター創刊にあたり

DE 研ニュースレター第一号 編集委員長 鬼塚 真
NTT サイバースペース研究所

1. ニュースレター発行までの経緯について

2008 年度よりデータベース学会、情報処理学会の DBS 研、そして電子情報通信学会の DE 研の一体運営が始まり、これを機に「DE 研の会員に対してよりよいサービスを提供できないか」ということを、DE 研の専門委員長である川越先生を中心に DE 研幹事団で検討する機会がありました。そこで「SIGMOD record のように一線の研究者のインタビューができないか」という旨を提案したところ、快く採用していただき、この第一回の DE 研ニュースレター発行の運びとなりました。

自分がかねてから SIGMOD record の distinguished database researcher の記事を愛読していて、そこには、経験のある研究者の研究や教育に関する体験談、その研究者の研究に対する取り組み姿勢、そして時には人生に対する考え方があり、非常に学ぶことが多い貴重な記事だと感じています。記憶に残っている言葉をいくつか引用してみます。

Jeffrey D. Ullman の若手に対するアドバイス。

You have to decide what game you are playing. What are you going for? Are you going for reputation, are you going for dollars, are you going for happiness?

<http://www.sigmod.org/sigmod/record/issues/0109/ullman2.pdf>

Jennifer Widom の若手に対する国際会議に通らなかったときのアドバイス(斉藤先生のインタビューにも出てきます)

My advice is to get that fire going, hit the job with a passion, and don't get discouraged.

<http://www.sigmod.org/sigmod/record/issues/0609/p57-column-winslett.pdf>

ひるがえって日本を見ますと、いろいろな学会誌もあり、またこれまでの DBWS、DEWS のようにコミュニティは非常に活発ですが、SIGMOD record のインタビューのように内容が濃くて、人に影響を与えるような記事がなかったと思います。そこで日本での著名な研究者の方の体験談をインタビューして DB コミュニティで共有したいと考えました。



2. 第一回ニュースレターのトピック

この第一号のニュースレターでは、日本だけでなくアジアを代表するデータベースの研究者である喜連川先生のインタビューと、もうすぐ開催される SIGMOD2008 に「Relational-Style XML Query」というタイトルで採録された斉藤先生のインタビューを取り上げることができました。また、SIGMOD2008 で Program Committee をされた NTT サイバーソリューション研の小杉さんからは、「SIGMOD2008 の Program Committee の経験を通して」という題で、今年の SIGMOD での査読の様子など貴重な経験談を頂くことができました。

喜連川先生のインタビューでは、連絡がなかなかとれず、インタビューは無理かなと思っていたら、突然連絡が来てインタビューをすることができました。「大学での研究の目的は何か?」、そして「本当の教育とは何か?」という点について、喜連川先生の深い考えと、感動的な経験について教えていただくことができました。また、斉藤先生のインタビューでは、斉藤先生の目指している研究者像をはじめとして、Relational-Style XML Query の研究の成功談と失敗談、バイオ系の研究とデータベース系の研究の対比などとても参考になる話を聞くことができました。

3. ニュースレターの今後の予定

ニュースレターは今後 3 か月に一回を目安に発行していく予定です。インタビューの他にも、大学の先端的な研究室紹介や、企業が得意とするプログラム開発の難しさや重要性に関する紹介や、国際会議の出張レポートなど、共有して価値のある情報を広く取り上げたいと考えております。掲載する記事の募集も行いますので、掲載したい記事の提案がある方は DE 研スタッフ de-staff@mail.ieice.org までご連絡ください。良い内容については積極的に取り上げたいと考えております。



喜連川先生インタビュー

喜連川 優

東京大学生産技術研究所

インタビュワー 鬼塚 真



「情報爆発時代に向けた新しいIT 基盤技術の研究」の領域代表者であり、またアジアを代表するデータベース研究者である喜連川先生にインタビューをしました。「大学での研究の目的は何か?」、そして「本当の教育とは何か?」という点について、喜連川先生の深い考えと、感動的な経験について教えていただくことができました。

以下の構成はメールアドレスのやり取りにインタビュー部分を補完した形式になっています。

1. 研究に関する成功談について教えていただけませんか?

- 並列データベースについて

並列データベースをなぜ開始したかといいますと、当時東大元岡研にいましたが、博士課程の学生は一人で1マシン研究するというという雰囲気がありました。パターン認識画像プロセッサ、プロトコルオフロード処理マシン、推論マシン、データフローマシンなど多様なプロセッサを研究する研究室でした。たまたまですが、小生は高級言語マシンからデータベースマシンはどうかと教授から示唆を頂きまして、ちょうどデータベースが注目されていたこともあり、データベースマシンに興味を持ちました。

一方データベース界は、小生が学生でありました70年代の終わりから80年代の日本では、穂鷹先生、増永先生、有澤先生などの多くの立派な先生方によるモデル研究が盛んでした。上林先生は、設計や問合せ処理手法などに関し広く理論研究をなされておられました。日本ではシステムの研究は当時富士通後に九州大学教授になられました牧之内先生の開発が注目されていましたし、又、データベースマシンは通産省のパターン大プロという国家プロジェクトで植村先生が指揮されておられました。一方、大学では本格的な所謂システム技術の研究は少なかった為、データベースマシンも含め、データベースのシステム技術の研究に取り組もうと思った次第です。当時は情報関連の研究者の数は極めて少なく、日本は同じ分野に沢山の人間をかけるというほどの余裕がありませんでしたから、人がやっていないところを探してやるというのが暗黙の原則であったと感じます。



当時、電総研の植村先生のお部屋に伺い、関連する資料をご指導頂いたことを今でも記憶しております。

- 成功した秘訣は？

成功したのかどうかは判りません。何をもって成功かも判りませんが(*)、自分なりに楽しい研究生活を多くのスタッフや学生に支えられて送られて来たと感じています。成功談という言葉から察しますと、カッコ良いサクセスストーリーのようなものがお話出来ればいいのでしょうかけれども、残念ながらそのような華麗な話は無いのが実情です。

(*) インタビューで、喜連川先生は「研究に関する成功の定義は非常に難しい」と繰り返されていました。

ただ研究の成功は「最終的に自分が満足する」こと無しにはあり得ないと感じています。他人の評価以前に自分自身の判断が大切です。そういう意味では、東大生研という、小講座制では無く研究室制の大変自由な場所で、ある意味で研究以外には何もすることが無い場所で、好き放題させて頂けたことは幸せでしたし、hash join 手法とそれに付随する種々の最適化に関する研究、hash join を実現する種々のシステム、即ち、ディスク一台と4つのバス結合マイクロプロセッサシステムから成る大変小規模な「機能ディスク」(*1)、機能ディスクを8台並べたスーパーデータベースコンピュータSDC(*2)、100台のパソコンをATMスイッチで結合したNEDO-100(*3)等多くのシステムを実装し、VLDB, ICDE, SIGMODなどに沢山論文も書けました。又、ハードウェアソータの開発もロジックレベルから進め、ボード、VLSI化まで15年以上かかりましたが、最終的に、Jim Gray sort benchmark で world record も出せました。この過程で三菱の伏見さんにはとてもお世話になりました。最近では、豊田准教授や田村さん、特任助教の鍛冶さんと共に socio sense と名付けたアジア最大のウェブアーカイブに基づく Web マイニングシステムを構築することが出来、また、特任助教の合田さんと storage fusion という新しい storage system の研究を進めております。これらのほとんどすべての研究のプロセスを中野さんがしっかりと支えてくれてきました。又、何事にも頼りない小生ですが、大変多くの秘書さんに面倒をみて頂きました。成功かどうかは判りませんし、非常に大変な時代もありましたが、大きな後悔はありませんし、全体を見ますと有意義な研究生活を送れたと感じます。

(*1)日本の科研費には3回不採択となり結局諦めました。当時のディスクは、大学で頂く研究費で一台購入すると殆ど他にはお金が残らないほど高価でした。scsi インタフェースの勉強が懐かしい次第です。68000系CPUで組んだディスク直結型データベースエンジンで、動的クラスタリングをDMA機構と共に、完全にハードウェアで実現したところが大きな特徴です。

(*2)その開発では富士通(株)に大変お世話になりました。

(*3)1996年頃のNEDO委託プロジェクトで、100台のゲートウェイ社製パソコンをATMネットワークで組んだシステムで、100台を動かすのに非常に苦労しました。私どもの研究室における並列データベースの実装技術については90年代で完結したと考えています。



成功の秘訣といえるものなどありませんが、あえて言いますと、多くのすばらしい先輩方に育てて頂き、また学生時代の良き仲間の伏見さんや、中野さんなどの良き同僚に恵まれたことが幸運だったのではないかと感じています。

卒業後、すぐに、指導教官(故元岡達教授)が急逝されまして、その後は、田中英彦先生とともに、故上林先生(京大)に非常にお世話になりました。上林先生には、海外で、色々な先生をご紹介頂いたり、研究のアプローチについて種々コメントを頂いたりしました。京大、東大の差別なく、自由かつ忌憚のない厳しいコメントを多く頂いたことは大変感謝しております。VLDBのプログラム委員会にも一緒に出させて頂き、そもそもどんな議論がなされているのか、どういうプロセスで論文の採否が決まるのかなど、とても若い時代に勉強させてもらいました。

加えて、小生は故高木幹雄先生に東大生研で拾って頂きましたが、海外に行くすべは高木先生に教えて頂きました。最初は、1年の2ヶ月半くらいは海外をほっつきまわっていたと思います。当時は、国際会議の論文を書いて博士号をとる人もおらず、小生も博士を卒業して初めて海外に行きました。米国の先生はどのような環境で研究しているのか、とても新鮮でした。いまでも Jim Gray のオフィスで議論をして頂いた時のことは覚えています。ちょっと話をし始めると電話が次々に掛ってくるのですが、じっと待っていますと、電話からコードを引き抜いて、「さあ、これから君のための時間だ」と時間を割いて下さいました。誰の紹介があるわけでもなく、自分でアポイントをとって行ったにも関わらず、若い研究者に丁寧に対応して下さいましたのには感動しました。もちろんそんな良い経験ばかりではありません。「オレはジャップとは口をきかない」と真正面から言われたこともありました。「what did you bring here?」と聞かれもしました。そうした色々な経験が研究者を作り上げて行くのだと感じます。

[インタビュー部分] 最初のうちは海外の論文を発表するという価値観は必ずしも良く分らなかったのですが、上林先生がトップレベルの国際会議に日本が出ることが重要だと強くおっしゃられ、「高度データベース」という重点領域研究をやらせておられた際、論文にウエイト係数をつけられまして、VLDB や SIGMOD はべらぼうに高い 20 点をつけて、「ここを狙うんです」ということを相当強く若い人に向かっておっしゃられました。「トップレベルの会議に出さないといけないんだな」と多くの研究者が感じたと思います。頑張るべき目標を定めて頂きました。データベース分野はやはり他の分野とは異なり、ジャーナルよりもこのようなトップコンファレンスが重要視され、小生も、ある時期まではほぼそれだけを目的に、論文を沢山書いてきました。だからと言って、それが成功かどうかと言われますと、それはどうか良く分かりません。価値観は時代とともに変わると感じます。

・ ビジネス的に並列 DB はどうだったか？

並列DBは オラクル 11g までも継続的に組み込まれており、既に並列 DB 機構は、commodity 化しているとも見做せ、ビジネス的には十分成功でしょう。並列処理は一般になかなか難しい領域ですが、最も成功したのは、データベースかもしれませぬ。ビジネス的には躊躇なく大成功と言えます。



2. メジャーな国際会議に論文を通すことについて

[インタビュー部分] デフォルトはリジェクトなんですよ。スタンフォードでもバークレイでもウィスコンシンでも 3 回ぐらいリジェクトは当たり前なんです。そういう中で論文をインプルーブする作業を彼らはやっている。それが当たり前なんだという感覚で論文を読むと文章の真意が分かります。こういう感覚がないと **passion** を維持することは難しいかもしれません。うちにも、1 回リジェクトされると非常に悲観的になる学生もいるのですが、困ったものです。リジェクトが当たり前だという環境になるべく早くなじむ必要があります。まず何でもいいからとにかくどんどん出し続ける必要があります。そうすると査読者からどういうリアクションが返ってくるかが分かって、何をなおすべきかが掴め、そして、一つでも論文が通ると、今度は海外の友人ができるようになって、情報交換をするに従ってどれぐらいの仕事量が論文としての世界的な標準かということがだんだん分かってくるんです。

僕は若いころに上林先生に育ててもらったので、ある意味ラッキーだったかもしれません。当時の VLDB のプログラム委員会は現在のウェブ会議とは違ってすべてリアル会議で、委員長が 1cm ぐらいの点数表の紙のコピー冊子を全員に配って、それを見ながら議論をします。そうするとどういう論文が通るのかというのが一目瞭然わかってくるわけです。どうして自分はこういうものが書けないのか？ 委員会の場では各論文に対するコメントも見ることができて、いい点をとっている論文に対するコメントも読めるわけです。そうするとどういう文章を書けばこういうコメントをもらえるかというリバーエンジニアリングができるわけです。なので、うちの研究室では、年間で 100 本以上の査読はゆうにしていると思いますが、学生に査読をさせるんです。なぜ査読をさせるかという、査読をする時にどういう観点から査読をするかを学生に体得させるためです。そうすると書き方のプロセスが逆転してくるんですね。査読する人間の立場を意識して論文を書けるようになるわけです。こういうことを少しずつ繰り返しやっていくことで、ポジティブフィードバックをかけてゆくわけです。

論文の査読結果に **reject** とか **accept** を書くのはものすごく難しいんです。neutral に丸をつけるなんて誰でも出来ます。Weak reject/weak accept も誰でもできる。でも、この論文は **reject** なんだということを書くにはものすごくシャープなコメントを書かないといけない。こういうことを若い時にするのはとても良いトレーニングになります。基本的に自分のところに回ってくる論文は自分のコンペイターの論文なんです。いままで自分が **reject** されてきた論理を当てはめて、相手の論文を見るようになります。逆に、相手が自分の論文を **reject** にしている時には、実は相手もよく考えているんです。Accept も主張するのは同様にとっても大変です。Sigmod に通してもらったとき、その一人の査読者がオレはお前の shepherd になったと書かれて査読結果が戻ってきました。涙が出るほど丁寧に論文を良く読んでくれていて、とても親切に指導をしてくれました。こんな人が居るのだということが判り、真剣に **accept** を出すことの重要性も感じます。そういうのを色々経験してゆきますと、ぐうの音もでない論文を書く方法がなんとなくわかるようになってきます。今は忙しすぎて丁寧に論文を書く時間が取れなくてとても残念です。

若い海外の研究者がわざわざ何度か研究室に来て、自分のやっていることを紹介してくれます。それ自体はとても面白いのですが、ある時、そもそも、どうして、うちに話に来ると素朴な疑問を投げて



みましたら、「あなたの感想が聞きたいから」というのです。そうやって、分野の近い色々な人のところで議論をふっかけて、どういう問題意識を自分のトピクスに対して持つのかということと事前に調査しているわけで、しっかりした論文を書いている人は実は多くの努力をしておられます。

さて、もう一つ大切なことは、研究者の仕事はスポーツと同じでいかに速くやるかでしか評価されないということです。アメリカのテニスは、5年か6年かで何本論文が書けるかが勝負です。この期間でどれくらい集中できるかで決まります。イタレーションのスピードが重要です。

しかし、一方で、国際会議に最適化した論文を書くことだけではむなしくもなりますので ほどほどにしましょう。

CIDER はそういう背景から作られた会議です。又、Top conference のTOT(test of ten years, 10年チェック論文選定委員会、10年後に振り返ってみて、インパクトのある論文を選択するもの)の仕事をしたことがあります。今はやりの citation で調べると驚くほどに多くの論文は引用されていません。つまり、良い国際会議に採択されても、それ自体大変なことなのですが、インパクトという意味では、まだまだ、距離があることがわかりましたし、google scholar、cite seer、citation index で全然スコアが異なることも明確になり、唖然としたことがあります。やはり、(言うのは簡単で、現実にはとても大変ですが) 本質を見据えた研究が望まれるところです。

3. これまでのご自身の論文の中で好きな論文を選ぶとすればどの論文でしょうか？

思い出の大きな論文はたくさんあります。VLDB, ICDE, SIGMOD で約20本の論文を書きましたが、なんといいですか、思い出が深いものも多く、多様な感情があります。

Application of Hash to Data Base Machine and Its Architecture

M Kitsuregawa, H Tanaka, T Moto-Oka - New Generation Computing, 1983

(ハッシュジョインの可能性を探る)

Hash-Partitioned Join Method Using Dynamic Destaging Strategy

M Nakayama, M Kitsuregawa, M Takagi VLDB 1988

(動的GRACEハッシュジョイン：ソフトウェアによる最適なハッシュジョイン実装法、東大基盤センター中山先生の力作です。ベンダは、なかなか明示はしてくれませんが、真相は定かではありませんが、米国の研究者の sigmod の論文に記載されている表現では、この手法は現在のDBベンダーのハッシュジョインの実装のコアになっているとされています。)

Design and Implementation of High Speed Pipeline Merge Sorter with Run Length Tuning Mechanism

M Kitsuregawa, W Yang, T Suzuki, M Takagi IWDM 1988

(ラン長動的最適化機構を組み込んだ超高速18段ハードウェアソータの実装、後にLSI化により商用化。精華大学楊先生と東京高専の鈴木先生が猛烈なエネルギーと長期徹夜で作られました。ある



時期、かなり多くの場所でご利用頂きましたが、プロセッサそのものが高速化され、最近はお番がなくなり残念です。100MBのDatamation Sort benchmarkで世界で始めて1秒を切りました。相当マニアックなので、ご存知の方はほとんどおられないでしょう。)

Query execution for large relations on functional disk system

M Kitsuregawa, M Nakano, M Takagi - ICDE 1989

(機能ディスクシステム：ディスク一個とマイクロプロセッサ4つでハッシュジョインを実現し、ウィスコンシンベンチマークで桁違いの高速性を実証。中野美由紀さんがIngresを移植し、徹夜でソフトデバッグしました。DeWittが日本に来た時にデモを見せたのを覚えています。)

Bucket spreading parallel hash: a new, robust, parallel hash join method for data skew in the super database computer(SDC)

M Kitsuregawa, Y Ogawa - VLDB, 1990

(バケット平坦化ジョインの論文ですが、リコーの小川さんが頑張ってくださいました。後にSDCではこのネットワークをハードウェアで実装しました。所謂多段スイッチングネットワークにおいて、スイッチに工夫を凝らしたものです。京都大学の富田先生から、アーキテクチャの講義の題材としてご利用頂いていると伺いました。)

Hot mirroring: a method of hiding parity update penalty and degradation during rebuilds for RAID5

K Mogi, M Kitsuregawa - SIGMOD 1996

(ホットミラー型ディスクアレイ、日立の茂木さんが巨大なシミュレータソフトを作られました。博士課程までで、vldbとsigmodにそれぞれ1本、計2つ論文を書いたのは茂木さんだけです。)

Parallel Database Processing on a 100 Node PC Cluster: Cases for Decision Support Query Processing and Data Mining

Tamura, T.; Oguchi, M.; Kitsuregawa, M. - Proc. of SuperComputing, ACM/IEEE 1997

(100ノードPCクラスタによる超並列データベース、データマイニング処理系。関連ルールマイニングの超並列化に試みた論文。SC:suprecomputingはHPC分野ではtop conferenceですが、三菱の田村さんとお茶の水大学の小口先生が当時はまだ筐体の大きい100個のパソコンを部屋一杯に並べ悪戦苦闘されました。100ノードのPCクラスタの構築と意味のあるアプリを96年の段階で稼働させていたのは、我々のグループが最も早いチームの一つであったと思います。)

Parallel mining algorithms for generalized association rules with classification hierarchy

T Shintani, M Kitsuregawa - SIGMOD 1998



(一般化相関ルールマイニングのハッシュ並列アプリアリ手法。日立の新谷さんと一緒に論文を書きました。これは上林先生の重点領域研究の期間に発表出来た貴重な **sigmoid** 論文です。)

機能ディスクシステム：関係データベース処理とその性能評価

喜連川優・中野美由紀 電子情報通信学会和文論文誌 1991

(機能ディスクというシステムの設計思想と、その実装について纏めています。)

もともとっと思いの深い論文が沢山あります。**GRACE** の **VLDB** へのデビュー論文は伏見さんと書きました。これは日本で **VLDB** が開催されたときでした。富士通の原田さんと書いた **KD** ジョインの **VLDB** 論文、中野さんと書いたとてもマニアックな分散共有メモリマシン用ジョインの **ICDE** 論文、更に超マニアックな鈴木さんのポインタサイズリングの論文(海外企業からの問合せが最も多かった論文です)、平野さん、中村さんと書いた **SDC** の論文(並列データベースにおけるフロー制御の複雑さは実装してみないと到底判らない深みがあります)、富士通の田村さんと書いた並列ヘテロマイニングの **VLDB** 論文、工学院大の山口さん、お茶大の小口先生と書いた **iscsi** の論文などなどきりがありません。実は今、沢山書きたい論文がありますが、時間がとれません。

4. *Google* の *GFS/MapReduce* や *Amazon* の *dynamo* についてどう考えられますか？

Computer science は 繰り返しの歴史である場合が多いと思います。**MapReduce** は典型例かも知れません。Dewitt の [blog\(*\)](#) がそれを端的に指摘しています。

(*)David J. DeWitt と Michael Stonebraker による **MapReduce** に対する批判に関する [blog](#)。

<http://www.databasecolumn.com/2008/01/mapreduce-a-major-step-back.html>

[インタビュー部分] **MapReduce** は シングルステージのハッシュアグリゲーションという観点では同じだが(ここで、ハッシュアグリゲーションとは所謂ハッシュジョインを集計演算に利用したもので、ハッシュ演算により極めて大きな性能向上効果が出る)、10万台のノード数で、且つ安定して動かしているのが **Google** の独自の技術でしょう。パリで行われた **SIGMOD** での **Google** からのキーノート講演では「**TCP/IP** のプロテクションコードを通過してしまうようなエラーが出るのが **Google** の世界である」という話があり感動しましたが、要するにそういう桁が違う世界に入った。これは昔の **ENIAC** の真空管と同じで、1万本の真空管の 多くが壊れることが前提にあった。つまり並列データベースのハッシュアグリゲーションと **MapReduce** とは設計の前提が違うので、単純に比較をすることは意味がありません。**Blog** の記事が批判的に書かれているのは、議論を盛りたてるためにやっておられるのではないのでしょうか。

ただこういう議論を通して、もともとっ、今のなわばりとは違うトピクスを考えることをデータベース屋としてはプロモートしていくことが必要なんです。リレーショナルスキーマに入っていないようなデータを我々の手中に入れて、次のプラットフォームをどんなものを作っていくべきか、ということ議論していけばいい。Joinの研究をしていた80年代初頭、学生のころ、情報処理学会全国大会で発



表しましたら、そんなデカイリレーションのジョインのアルゴリズムなんて要らない、誰が必要としているのかなどと厳しいご意見を頂きました。当時、MapReduce が一日 20 ペタバイト処理するとは誰も考えなかったと思います。新しい領域を常に新陳代謝することはとても大切です。

5. インドやアフリカを訪れて人生観が変わったと聞いたのですが、どのようなことでしょうか？

極度に裕福な日本にいと、そして、常に、サンフランシスコやNYなど IT系の国際会議はどちらかという civilize されたところばかりで開催されますので、全く違う環境を経験することはとても重要だと思います。ict4b(*)などの発想がどこで生まれるのかという、より広い視点も必要だと感じます。

(*)Eric Brewer による、ICT 技術を生かして発展途上国を支援するプロジェクトでの取り組みの一つ。

http://www.citris-uc.org/research/projects/ict4b_a_scalable_enabling_it_infrastructure_for_developing_regions

[インタビュー部分] ict4b のように発展途上国を支援する技術を研究するという事は、日本に住んでいる限り思いつかないでしょうね。コンピュータサイエンスの国際会議は、パリやサンフランシスコやニューヨークで行われることが多いので、プライベートの休暇でインドやアフリカに行ってみて、確かにこんな世界があって、この世界に対して我々が何ができるかを考えるチャンスは、自分のウィルを持ってやらないと、絶対に想像ができないですよ。人生観が変わるといのは、言葉では説明が難しく、一回行かないと分からないので、行く価値はありますよ。

中国があれだけ不安定なのに、なんでインドがこれだけ安定しているかも非常に不思議ですよ。なんでカースト制度があって安定しているのか。インドは都市部では、カーストのレイアを超えて結婚が出てきています(地方では駄目ですが)。インドの農業は日本と似ていて、農業が非常に細分化されているんですね。一人の農民が持つ土地は非常に狭いんです。一日平均 30 人の自殺者が出るそうです。それは、成長率が良いといわれて高いローンを組んで高価な種を買う。でもその種はうまいこと育てたら成長するんだけど、インドの劣悪な環境では水を丁寧にやる、肥料を丁寧にやる、殺虫剤を丁寧にやればできるんだけど、そういう管理技術がそろっていないので失敗してしまう。そうすると負債ばかり背負って、自殺が起こってしまう。

こういう問題に対して、うちのインドからの留学生だったポスドクが、最近テレビに出るくらい有名になっているんですが、人力 IT でこの問題を解決しようとしています。サーチエンジンでいうところの「人力リサーチ」のように、要するにインドでの人件費が限りなく低いことを利用して、IT でやるところの相当のところを人手でやるセンスの良い方法を生み出しました。仲介屋を安い賃金で雇いまして村々を訪れて、栽培している作物の様子を写真の写真を撮って、町のセンターまで行ってそこからデリーまで通信して送る。そうするとデリーにいる農学の大学の教授が写真に写った葉の上の虫、枯れ具合を見て、どういう農薬をどれくらいいつ使えばいいかを細かく指示するんです。この指示を先の仲介屋が丁寧に字の読めない農民に口で伝えます。この準リアルタイムシステムで、収穫量が飛躍的に上がったそうです。最近では、アフリカの国もこれに注目して招待されていると聞きました。このシ



ステムが sigmoid の論文のネタになるかという論文にはならない。けれども彼(インドからの留学生)がうちの研究室にいて言ったのは「僕はとにかく自分の国の為になることをやりたい」。こういう感覚は(インドに)行かなきゃ分からない。話を持ちかけても、最初は誰も相手にしてくれなかったようで、けれども彼はこつこつ2年ぐらいかけて取り組んで、やっと動き出した。そしたら政府もやっと認めてくれて、大きな予算をくれたそうです。その学生は、今はハイデラバードの full professor になっているんですが、去年研究室にインドの自分の学生を連れて来て、その学生に向かって「これを見て。僕はね、この研究室のこの机で4年間勉強をしたんだよ」と言っていました。こういう純粋な気持ちのポストドクが自分の国に戻って、頑張ってくれるのを見ると学校の先生冥利に尽きますね。近々、stockholm challenge awards を受賞するそうです。

「論文を書くことが価値じゃない」。私(先生)はポストドクに対して「最終的に何をやらないといけないのか、じっくり考えましょう」と最初に言うんです。「学生と君らは違うんです。学生は大学にとってみるとサービスを受ける側で、教官やポストドクはサービスをする側なんです。君たちの給料はどこから来ているかを振り返ってみましょう。それは日本国民の血税から来ているんだから、最終的には税金を払っている人のことを考えるということが妥当なのではないか。ちょっとカッコ良すぎるかもしれませんが、つきつめると研究をするということは納税者に対して社会価値を返すことと思う。僕たちは納税者から聞かれたら、何をしようとしているか、何をしてきたかを判りやすく言えるようにしないとイケない。論文を書くということはあくまでパイプロダクト。ファーストプライオリティじゃないと思う。」というようなことを話して学生と議論しています。先ほど「成功」については、自分が満足出来るかどうか肝だと申しました。これは一見、個としての研究者の満足感を重視し、ここへ来て、社会還元ということいいますと大きく矛盾することを言っているように感じられるかも知れませんが、小生の説明能力が極めて低いことを露呈しているかと思いますが、研究の方向性、目標というようなことは、胸を張って、そして情熱を持って、社会還元を意識すべきだと考えています。社会還元には色々な方法があると考えます。もちろん世界がアツと言う論文を書くことは大切でしょう。申し上げたかったことは、自分の成果を表現する際に、論文が採択されたということを見せることがとても判り易い表現であるが故に、ともすれば、論文をだすことだけが目的になってしまうことがあり、そうではないということを自戒も込めて申し上げたかったのです。大前提として、その研究成果を自分が愛せるかどうか大切と感じています。そしてまた、論文だけが、自分の社会還元に対するエモーションを表現するものではないこともお伝えしたいのです。先の留学生は、インドに戻って、大学人として単に論文を書くだけではなく、母国に大きな貢献をしてくれてうれしく感じます。もちろん、競争力のある技術を生み出して、それが製品化され国に多く納税するという企業価値につながってもOKなのですが、それ以外の貢献の仕方もあると。でもこういう考えが正しいのかどうか判りませんし、又、これから、考えが変わるかも知れません。



6. グルメであると聞くのですが、よくご利用される場所やお勧めの食べものなどありますか？

それは 長い間、六本木にいただけで、今は駒場キャンパスという食の選択肢が限られた場所ですから。大阪生まれですから、食は重要です。逆に言いますと、それ以外はあまり興味がありません。

7. 近年、DB 分野の研究と実用の乖離が進みつつあるように思われますが、これについてどうお考えでしょうか？

そういうことはないと思います。多くの研究者が（小生の予想以上に多くの研究者が）XMLの研究をしています。産業界はXML関連の開発を熱く進めており、そういう意味では、そんなに乖離していない。とりわけ、DB が成功してきたのは、ニーズドリブンに進めてきたからでしょう。これほどデータベースソフトウェア関連の市場が大きく維持出来ているのは、情報関連の分野でも珍しいと思います。XMLに限らず、データウェアハウス、OLAP、association を起点とするデータマニングなどすべて、インダストリが問題を規定してきた経緯を見つめなくてはなりません。

一般に、研究と実用の距離は常にむずかしい問題で、実用ばかりみているといわゆる disruption は起こりません。一方、ある程度現実を見据えないと デルタ論文ばかりになってしまいます。特効薬はないと思います。人材育成がとても大切だと感じます。これも言い出すときりがありませんが、「やりたいこと」と「やらないといけないこと」は必ずしも一致しません。今のITは後者からもっと攻めて行くべきだと思っています。西海岸はそこをうまく回していると思います。小生の主張は10以上前から一貫しておりまして、企業と大学はもっともっと「問題を共有」すべきです。よく日本の大学は情報産業に何も貢献していないという言われ方をしますが、このあたりがネックだと感じています。そもそも、日本のソフトウェア産業自体が弱いわけですが、問題の共有によって、改善されると強く感じています。

8. 今後の有望な研究課題や取り組むべき/取り組む予定の課題などあれば、いくつか教えてください。まず自分の好きな研究者をえらび、その動きを見て、その先を予想するというところから始めるのが良いと思います。そうでないと、今聞いた課題はもう遅すぎるからです。

9. 後進の研究者へのアドバイスを頂けませんか？

何を研究テーマとするかという点では、我々の研究室に来た学生に必ず言うようにしていますのは、「自分がやりたいこと（興味があること）と、自分が他人より勝てる（自分の得意な）ことのANDをとって、そこから攻めなさい」ということです。なるべく早い段階から、研究テーマを自分でしっかりと見つけてゆく能力を身に付けることが大切と感じ、このように言っています。とりわけ大学の研究者になりたい人は、IT は動きがありますから、ずっと同じことを一生やるということはまず稀だと思いますので、どんどんテーマを探してゆく必要があります。その際のヒントです。

又、研究は当たったり当たらなかつたりいつも良いテーマに当たるとは限りません。ですので、色々な興味をもっておくのが良いと感じます。



論文という点では、どんなトピクスでもいいので、一度は **top conference** で発表してプレゼンスを得ることが大切と感じます（これは上林先生のお考えと同じです）。自分がどれだけのエネルギーを投入するとWWレベルに達するかを体得することが大切だと感じます。一度もやらないと、どれくらい努力すればいいかという感覚が養えません。

ただ、一方でそれだけを目指することは不適當です。SONY CSLの所先生は一つの新しい学問領域の創出を目指すべきであるとおっしゃられます。まさにそのとおりだと感じます。つまり、どの国際会議という場も決まっていなくていい、全然違う、大きな領域に挑むことも重要です。が、こんな大きなことは小生のような小さい器ではちょっと難しすぎますが。

研究だけではなく自分の活動の場である学会へのサービスも大切です。米国のテニユア評価には **service** の欄があります。研究だけはダメです。30代後半は通信学会和文論文誌の編集委員をさせて頂いておりましたが、増永先生に委員長になって頂いて「オブジェクト指向」の特集号で、厚さ1センチを実現し、厚さ1センチのステーキ打ち上げを六本木でしたことがとても良い思い出です。40過ぎの時に通信学会データ工学研究会の委員長を仰せつかりました。前委員長の西尾先生より、DEWS（当時）を大きなデータベース研究者の集いにしてゆくことはとても大切と申し送りを頂き、当時、リアルセッションからパラレルセッションにするかどうかで、大きな議論をしたのを覚えています。情報処理学会との連携について当時DBSの委員長の田中克己先生に大変お世話になりました。その後、**sigmod** の日本支部長や情報処理学会の理事を仰せつかりました。自分の研究発表を支えてくれているフレームワークに対して「適切なエネルギーで」奉仕することは必須で、国際会議でもそうですが、そこから信頼の輪が出来て行きます。やり過ぎも良くありません。断る勇気も必要です。しかし、何もやらないというのもダメで、バランスが必要です。自分の論文が採択されるとすると、その背後には、それを読んでくれている人が何人もいて、国際会議で発表すると、その会場のアレンジからプロシーディングスの作成から、海外の **registration** のビザの処理まで多大な労力を払ってくれている「誰か」が背後にいることを認識しなくては行けません。どんな鈍感な人でも、何回か会議に出れば、それは感じるはずで、そういう努力を自分もちゃんと奉仕しないとアカデミアの世界は回って行かないことを体感して欲しいと思います。上林先生に **VLDB Trustee** にご推薦頂きまして、6年間、毎回の会合に参加しました。メンバの中では、最年少でたいした貢献も出来ず、むしろ勉強させて頂いた次第でしたが、一年に一回とは言え、朝から晩までみっちり非常に丁寧に **endowment** 会合をします。topレベルの会議をtopレベルに維持するのは裏で大変な作業がなされていることを肌で感じました。ICDEのステアリング委員会でも同じですが、会議に慣れていない国での開催にはとても丁寧に、その国情に合わせた配慮が議論されます。

ある早朝、故上林先生から「今、犬と散歩しているんだけどね、XXXのことを考えているんだ。どう思う？」と携帯電話から連絡を頂いたことがあります。とても眠かったのですが、こうやって若手を育てて下さっているのだなあと感じました。植村先生や増永先生からは、年賀状をご覧になられて「いつ勝手に結婚したの？」と温かい電話を正月早々に頂戴し励まして頂きました。年配になるとこうやって若手をエンカレッジしないと行けないのだと感じました。



本インタビューの母体である電子情報通信学会という観点では、H19年度に、一年間東京支部長なるものを致しましたが、自分では普通なつもりなのですが、当初、評議員の先生方から「今までにない、unusualな支部長。こんな変なことを言う人は見たことが無い」などと言われてまして、ただ、最後には「変な人だと思っていましたが、やっていたら楽しかったです」と感想を言って頂きました。どうも小生はまだ一人前ではない世間知らずの子供のようでアドバイスなど到底おぼつかないのですが、何かありましたら、ろくな意見はもらえないという前提で気軽にお声をかけて下さい。50肩になって、老眼もひどく、あまりたいした能力も無いのですが、これからは、もっと積極的に若い方々のお手伝いをしないといけないと感じています。この年になっても色々不慣れなため、忙しくしておりますが、時間はどこかで工面したいと思います（このインタビューもとても遅れて失礼致しました）。厭という程のrejectとそこから這い上がる経験もありますし、何度も胃カメラを飲みつつ研究をしてきましたので、ほんの少しは何かお役にたてこともあるかもしれません。

最後に、そもそも、25年前と今では、計算機科学分野の状況はすさまじい進歩で、大きく変わって来ました。ですから、年寄りの言うことなどに耳を傾ける必要など無く、細かいことに気を捉われず、「癒しのある家庭と良かったなと思える温かい人生」を如何に実現するかをお考えください。研究以前に人生の最適化がとても大切です。



斉藤先生インタビュー：本当に「使える」XML を目指して

斉藤 太郎

東京大学大学院 新領域創成科学研究科 情報生命科学専攻

インタビュワー 鬼塚 真



年々国際会議のレベルが上がっている中、「Relational-Style XML Query」というタイトルで SIGMOD2008 に採録が決まった斉藤先生にインタビューしました。

斉藤先生の目指している研究者像をはじめとして、Relational-Style XML Query の研究の成功談と失敗談、オープンソースとして公開されている Xerial、バイオ系の研究とデータベース系の研究の対比、そして PhD の学生へのメッセージについて聞いてみました。

1. 今回の SIGMOD に代表されるような研究だけではなく、Xerial をはじめた皆さんのツールも開発されていますが、斎藤さんが目指す理想像(研究者像)や研究のスタイルを教えてください。

xerial.org で公開しているツールは、プログラミングを快適にするために作ったものがほとんどです。研究テーマに関連する道具を作っているうちに、各プロジェクトで共通に使えるライブラリがそろってきたという様子です。また、しっかりとしたライブラリを作りたいという思いと、研究に必要な部分に絞ることの間に、自分の中で葛藤があります。プログラムをすること自体は楽しいのですが、それだけしていると研究としてまとまらないことが多いのでバランスが難しいのですが、プログラミングと研究を交互に行うように自分のペースを探していかなければならないと思っています。

理想とする研究者像は、役に立つものを自分自身の手で考え出して、実際に作れる人、ですね。但し、ただ作るのではなくて、技術を活かしてクオリティの高いものを作り上げる職人気質も持っていて、なおかつ、こんなものがあったら便利になるとか、研究として新しいものを作る、という想像力を強く働かせるデザイナーでもあったり。Xerial(エクセリアル)プロジェクトも、DBMS を自分の手で作り上げたいという思いから始まっています。新しくいいものを作り上げたいというのが大前提にある気がします。

研究者のスキルとして必要だと考えているのは、アイデアを式や文章の形にまとめる能力。しっかりと定式化や論文ができていれば、実装するのは非常に楽になりますし、コードの中身を他人に理解してもらいやすくなります。今回、SIGMOD の成果が出たことで、協力者を募って、Xerial の開発を



ペースアップさせていきたいところです。

研究スタイルは、例えば DBMS なら、それを既存のシステムになるべく頼らず、一から自分で実装するという形を取っています。研究成果を出すという意味では、ずいぶん回り道をしていることが多いのですが、実装してみて初めて実感できる部分というのが、データベースには多いと感じています。トランザクション処理でスループットを上げる難しさ、クエリコンパイラの実装、ディスクレイアウトの扱い、なぜ既存の DBMS だと遅いクエリがあるのか、などなど、システムの内部構造まで知らない理解できない部分がありますし、そこに新しい研究テーマが眠っている場合もあります。そういった経験を積み重ねると、プログラマとしてできることの範囲も広がり、自分自身のライブラリも充実して開発効率も上がっていくので、研究がどんどん楽しくなってきますね。

サーベイをするときや、新しい研究テーマを決める、アイデアを練る段階になると、一切プログラミングをしない週、長いと月というのが出てきます。紙のノートと延々向きあったり、ネットワークから敢えて切断された環境に身をおいて、執筆に集中できるようにしたりします。プログラミングをしていないと、動くものがないという焦りがでてくるのですが、後々になって評価してみると、PC から離れて熟慮している時間から、いいアイデアや、ストーリーが生まれてくることが多いようです。ただ、実装もそれなりに時間がかかるものなので、どちらかに偏りすぎないように、いいバランスを取れるように今でも試行錯誤しています。

2. SIGMOD に採録された研究の概要について教えてください。

・ どのような問題を想定したか？

Relational-Style XML Query で扱っているのは、リレーショナルデータ、つまりテーブル形式のデータを XML から取り出す問題です。この問題を応用すると、XML という木構造を持った複雑なデータに対して簡単なシンタックスでクエリを実行する方法につながります。

・ その問題がなぜ難しかったのか？

テーブル形式のデータを、XML の木構造で表現しようとする、使える木構造の種類にかなりの自由度があります。これを「構造のゆらぎ」と呼んでいます。XML データを作るときには、使える木構造のパターンが多くて自由度が高いのに、クエリを実行するときには、XPath や XQuery では、パスを使って1つのパターンを指定する必要があって、データを作るときにある自由度をカバーできるだけの能力が今のクエリ手法にはないんです。このギャップ（データベースの用語だと impedance mismatch といいます）を乗り越えることが課題でした。

・ どのようなアイデアで問題を解決したのか？

XML データは確かに木構造をとっているのですが、その大部分は、本質的にはリレーショナル（つまりフラット）な構造に置き換えられると考えました。

従来の XML クエリのアプローチでは、「XML は木構造だから、木をたどるクエリを実行する」というとても素直な方法をとっていました。XPath や XQuery のようなパスによる問い合わせは、この分類です。



今回の研究では、それとは違った発想をし、もともとはフラットなリレーショナルデータがあつて、それを XML で表現するときに、たまたま木構造になったんだと、考えることにしました。そうすると、問い合わせするときには、ユーザーはリレーションを指定するだけで良く、後はシステム側が自動で、考え得る全ての木構造のパターンをカバーできるようになりました。

このアプローチそのものは、amoeba join というアルゴリズムで既に発表していたのですが、リレーションの中にリレーションが含まれるようなネストした構造を持つ XML データの場合だと、正しい木構造のパターンを amoeba join のみで見つけるのは、非常に難しかった。

そもそも、ユーザーがどんな意図をもってその XML データの木構造を組んだのかがわからないと、本当の正解が定まらない問題なんです。そこで、データ構造の持つ意味(semantics)を明確にするために、リレーションに加えて関数従属性(Functional Dependency, FD)を XML の世界に持ち込みました。

リレーションも FD も、リレーショナルデータベースの世界では、ユーザーはその存在を意識しないで使っているものです。たとえば、テーブルスキーマを決める、というのは、リレーションと、FD の集合を定めていることに等しい。Relational-Style XML Query では、XML におけるリレーションと FD、つまりテーブル構造データは、様々な木構造で表現してもいいという定義をしています。このゆるさを XML の世界に導入することで、リレーショナルデータベースと同じ感覚で、XML データを扱えるようになっていきます。

3. 今回の論文は問題設定が非常に鍵だと思いますが、どうやってその問題の発見に至ったのでしょうか？

実を言うと、最初からリレーションを XML から取り出す、という問題を想定していたわけではありません。むしろ、この問題設定は、自分が今まで研究してきたことを、データベース分野の人に一番良く理解してもらえる例として、捻り出したものです。

研究のきっかけとしてよく覚えているのは、バイオインフォマティクスの分野で SCMD (<http://scmd.gi.k.u-tokyo.ac.jp>) という Web データベースサーバーを作成していたときのことです。酵母菌の個々の遺伝子を破壊して細胞形態の変化を調べるための画像データベースだったのですが、中身は遺伝子ごとに分類される階層構造を持ったデータがほとんどでした。ちょうど僕が XML データベースの研究をしていたこともあって、研究室内で、階層に強い XML で SCMD のデータを表現してみようという話になったのですが、いざ、細かいデータを XML で表現しようとする、階層の親子関係を入れ替えて XML データを組みたいことが多々あって、木構造のスキーマを定めるということが、非現実的なことを実感しました。特にバイオ情報だと、最初から一つのスキーマセットがあるわけではなくて、後から解析してどんどんデータを追加していくことでスキーマも変化していくという状態で、その度に構造をかちっと決めて SAX や DOM を使ってプログラムを作ることに難しさを感じていました。それがきっかけで、データとして実質的な意味をもたない木構造が XML 中にあると考えはじめて、木構造の組み方が変わってもクエリできる amoeba join というアルゴリズムを考えました。



当時の研究ノートを振り返って見たのですが、2006年6月頃にメモしてあるアイデアを見ると、amoeba joinによる問い合わせができたから、次はupdateだと考えたようです（その当時の思考過程はまったく覚えていないのですがw）。updateを行う際には、XML中のどの位置を更新するかユニークに特定するために、どうしてもkeyの概念が必要になります。IDのようなものですね。階層を持っているデータのkeyについては、Peter BunemanたちのグループがKeys for XMLの論文で提唱していましたが、これは階層が入れ替わる状況に対応していなかったもので、ではamoebaの考え方を使得keyを定義したらどうなるだろう、というところが今回の論文のはじまりです。またkeyというのは、よくよく考えると、関数従属性(FD)の特別なケースであることに気がつきました。このようにupdateのためのkeyから始まってFDに至ったわけです。実は、昔XMLのupdateの研究を卒業研究していたので、昔やっていたことが生かされたということがありますね。

FD自体も最初から使っていたわけではありません。構造が入れ替わる階層キーを持ったXMLデータもしくは階層がぐちゃぐちゃな構造に対してはpath expressionが使えないので、どうクエリを実行したらいいかと考えていて、node間にone-to-manyなどのdependency(FD)が定義されているなら、その情報を使えば、amoeba joinを組み合わせて解けるという結論には、既に2006年6月の時点で至っていたようです。今回の研究の原型はかなり昔に出来上がっていたんですね。

Relational-Style XML Queryでは、本当に様々な問題にチャレンジしています。構造のゆらぎであったり、パスを使わないクエリであったり、XMLにおけるkey(keyはFDの一種)やdatabase integration、incremental updateにまで言及しています。どれを主題にしてもおかしくないのですが、Relation + FDという枠組みでいこうと考えた直接のきっかけは、僕の博士論文を審査していただいた国島先生が自身のブログで僕の研究を、「XMLに対して選択・射影演算を定義し…」と要約してくれていたことです。自分自身ではalgebraを定義しているつもりはなかったのですが、データベースの研究者にとって、relational algebraを使って書くと問題が明らかになるんだと実感し、最終的にrelationをXMLから取り出すという問題設定にしました。非常に紆余曲折があったわけです。

始まりは実用面から始まり、バイオ系で必要なものをどんどん考えていって、最後にデータベースの人に分かってもらえる問題設定にたどり着きました。

4. 論文というのは研究成果の氷山の一角だと思いますが、その裏にある失敗談や成功談などあれば教えてください。

失敗談といえば、今回のテーマは、VLDB2007にも投稿したのですが、そのときには落とされてしまいました。さして明確な根拠もなしにこの方法は便利だなどという書き方が多かったため、reviewerにあまりいい印象を与えなかったのでしょうね。一番痛かったのは、研究のmotivationがわかってもらえなかったこと。自分はこれだけアプリケーションがわかっているのに、reviewerには伝わってなかったのはショックでした。今読んでみると、書き方が下手だなと感じるので、落とされた理由がよくわかるのですが。



そのときから、テクニカルなところは何も変えないで、SIGMOD に向けて論文の文章だけ書きなおしたんです。締め切りまでの 5 ヶ月間（実際執筆していたのは締め切り前の 1 ヶ月半）で、タイトル、abstract から実験結果まで 論文は 9 割以上書き直しました。テクニカルなことはそのまま、ただ、ストーリーの書き方を変え、魅力的なアプリケーションを前半に据え、ロジックのおかしいところや、根拠が乏しい部分も、1 パラグラフごとに検討して、reviewer を mislead しないように丁寧に書き直していくという作業を、毎日少しずつ森下先生と一緒にやっていました。森下先生自身も、海外のチームと一緒に SIGMOD の論文を書いた時は、1 パラグラフごとにディスカッションして、毎日少しずつ直してという経験をしていたらしいです。向こうの研究者はそれくらいやっている。ネイティブスピーカーがそれくらいやっているのだから、英語に不慣れな日本人はもっとやらないと読みやすい論文はできないんだ、そういう経験談を聞いていたので、実際本当に自分でそれくらいやってみようという気になったのが大きかったですね。

不思議なことに、いつも厳しいことや、趣旨を勘違いして論文を落としてしまう人達だと思っていた匿名の Reviewer も、丁寧に論文を書いたときには、ちゃんと味方になってくれるんだと感じました。論文の不備の指摘の仕方も、今回は非常に穏やかでした。確かに、自分が review をする立場にたったとき、ぞんざいな書き方をされては、とてもいい評価は与えられないですよ。同じ研究者として、面白いと思った研究には、こうしたらもっとよくなると、素直にアドバイスを書けるのだと思います。

5. 齊藤さんはお子さんの送り迎えをされていると聞いていますが、毎日お子さんの面倒とか見ているのでしょうか？

学生結婚したので、修士の頃から子どもと一緒に研究をしているという様子です。学生だからできた子育てという感じがします。時間があって、保育園に朝子供を送ってから、研究室に来てプログラムを書いたり論文を書いたりしていました。今朝も子供を送ってきました。子供が生まれたばかりのころは大変でした。子育ても慣れないし、研究時間がガクンと減ってしまって、今までは夜にプログラミングを始めて夜中までやっていたものが、一切できなくなってしまいました。けれど、意外なことに、ちゃんと朝起きて子供の時間に合わせてという生活リズムができてくると、うまく自分の中で休みをとれるようになりました。今も土日は休んでリフレッシュしています。そして月曜日から研究をやりたいという気持ちになって一週間がスタートできるので、いい面があるんです。メリハリが付けられるようになったと思います。

6. 開発されているツールについて教えてください。

書きためていたコードを合わせると 20 万行ぐらいはあるんじゃないかと思います。修士の 2003 年ぐらいから書いています。相当書いていますね。SQLite が 7 万行で、PostgreSQL が 80 万行のようです。Xerial というサイトを運営していますが、まだ、ひとつの DBMS としてまとまった形では、何も提供していないので、残念ながら、ユーザーはほとんどいません。開発者もまだ僕 1 人です。ライブラリとしては利用価値のあるものが多いので、自分自身で使うことが多いのですが、一般の人に使っ



てもらふためには、もう少し、研究としてではなく、ユーザーのための開発や、ドキュメントの整備をする必要があると考えています。

7. 森下研究室はおもにバイオインフォマティクスの研究をされていますが、データベース系と対比的に、バイオ系の研究を通じて得られることがあると思いますがどのようなことがあるのでしょうか？

2007年の4月から森下研究室で助教をさせて頂いているのですが、こちらにきて驚いたことは、バイオインフォマティクスの研究者たちは、データベース研究者より、はるかにデータベースを現場で使っているし、その使いにくさを身をもって実感している人達なんです。フリーのDBMSは、数千万～億単位のレコードを挿入しようとすると、悲鳴を上げるけれど、彼らはもっと多くのデータを必要としている。毎週のようにテラ単位のデータを扱わなくてはいけない、など、SIGMODの実験結果でもお目にかかれなような規模のデータを日常的に扱っています。XMLの記述力を欲しているけれど、この規模のデータを扱えるシステムが身近にないので、結局RDBに落として格納している、などなど。ディスクの速度、通信するネットワーク速度、システムの使いやすさなど、いろいろな要素がボトルネックになることに気がきますし、彼らが幸せになれるDBMSを考えれば、それがそのまま良い研究テーマになるとも感じています。不便さを解消してあげなきゃと、常々思っています。

バイオインフォマティクスは、データベース屋さんにとって、アプリケーションや解決する問題には困らない場所だと思います。最近、Data spaces, Data provenance, Map-Reduce的な分散処理などの研究分野がデータベース分野で開拓されてきていますが、それを研究レベルではなくて、実用レベルで欲している人たちが正にここにいます。研究成果を出して終わりなのではなく、生物系の研究者に意味のある結果やサービスを提供するために、開発は常に実用やサービスにもっていくことが前提です。その意味で、バイオインフォマティクスは、純粋な情報系より、コーディングの能力が要求される分野です。僕自身も、既に去年の5倍以上の量のコード(10万行!!)を森下研に来てから書いていますが、まだ、それでも人手が足りない。腕に覚えありという方には、ぜひ飛び込んでほしい分野ですね。

8. PhDの学生へのメッセージを頂けませんか？

僕もまだ研究者として、駆け出しなのであまり大それたことは言えませんが、データベースの研究者を目指すなら、早いうちに、SIGMOD, VLDBなどの国際会議に参加してその雰囲気を感じておくといいと思います。研究テーマというのは、絶対的に存在するわけではなく、人の中、コミュニティから生まれていることがよくわかると思います。参加が難しいなら、公開されているKeynoteスライドのスピーチの動画を観たり、google videoを活用してプレゼンテーションのビデオを見るなり、今はいろいろな方法があります。学振などで自分の研究費が持てるなら、高価なマシンを買うより、旅に出て学会に参加することを薦めます。自分も実際そうしました。

論文を通して参加するのが一番ですが、国際会議への論文の投稿数が増大し、分野も拡大しているので、一昔前に比べてなかなか論文が通りにくくなっているという話をよく聞きます。そういうときにJennifer WidomがSIGMOD Recordのインタビューで言っていた、"Don't get discouraged (がっかり



しないで)" という言葉が励みになります。たとえ、いい論文でも、**reviewer** のめぐり合わせで、落とされることもあるからです。

最後に。プログラミングのスキルは自分の財産になるので是非身につけてほしいと思います。アルゴリズムをすぐに考え出すのは難しくても、考えたことをプログラムとしてスムーズに表現する能力は、経験でまかなえる部分です。言語の仕様やライブラリに慣れること、オブジェクト指向、デザインパターン、リファクタリング手法などを学んでおいて損はありません。いざ、プログラムを書くところで尻込みしてしまうようだと、できる研究の幅が狭まってしまいますし、常識を覆すような大きな仕事はできないと思います。DBMS を 1 から作ってみる、あるいは **Top Conference** の論文を読み込んで実装してみる、などの課題に取り組んでみると、文章を読む力とともに、実装力もついて一石二鳥です。



SIGMOD2008 の Program Committee の経験を通して

小杉 尚子

NTT サイバーソリューション研究所

小杉さんは特に音楽検索の分野で研究と開発を両方されていますが、SIGMOD2008 で Program Committee をされましたので、今年の SIGMOD での査読の様子などの貴重な経験談を頂きました。

SIGMOD2008 でプログラム委員をやらせて頂きました、NTTの小杉です。
皆様のお役にたてそうなトピックを選んで報告いたします。

皆様ご存知の通り、SIGMOD2008 の PC Chair はニューヨーク大学の Dennis Shasha 教授です。私は、2005.6-2006.8 の期間、客員研究員として Shasha 教授のところに海外赴任しました。そのご縁で、今回 SIGMOD2008 のプログラム委員にお声がけ頂きました。まず最初に、Shasha 教授について少し紹介させていただきます。

Shasha 教授は大学では主に数学、データベース、バイオインフォマティクスなどがご専門ですが、そのほかにも本の執筆 (<http://www.cs.nyu.edu/shasha/outofmind.html>) や子供の教育など、非常に広範囲にまた精力的にご活躍されています。日本では、2006年の4月まで日経サイエンスに「パズリング・アドベンチャー」を寄稿されていたので、そちらを御存知の方も多くいらっしゃると思います。

「教育」というのは、Shasha 教授にとって非常に重要なキーワードだと思います。身近に実感することができたのは、客員研究員として過ごさせて頂いた1年余りですが、Shasha 教授は非常に忙しい中でも、学生の指導や授業にはとても熱心だったのを覚えています。常に、個々の学生のレベルや志向を考慮して（ちょっと高め）テーマやサジェスションを与えて、しっかりサポートする（仕事を振りっぱなしにしない）という指導方針を貫いていらっしゃいました。Shasha 教授の下でのテクニカルな勉強ももちろん有意義でしたが、このような、「人を育てる」ということを直接拝見・体験させて頂く機会を得たことも非常に大きかったです。

今回の SIGMOD2008 の採否にも、Shasha 教授の「人を育てる」という姿勢が非常に強く現れていると思います。その意向を汲んで、少しでも皆様の今後の研究活動の参考になることがあればと思い、SIGMOD2008 のプログラム委員としての活動を通して得られたことをここにまとめさせて頂きたいと思います。



プログラム委員の実質的な活動が始まったのは 2007.11.14 です。abstract submission の deadline が過ぎたので、

11.23 までに査読したい paper の bidding を済ませるようにとの指示でした。

SIGMOD 2008 への submission は全部で 625 本で、去年の約 30%増とのこと。

活動が終わったのは 2008.2.14 で、同日 23:30(PST)以降は査読結果を変更しないようにとの指示がありました。

活動が始まる前/最中は、再三にわたって Shasha 教授から査読指針の説明がありました。骨子は以下です。

1. 査読にあたってはそのペーパーの良い部分を見つけるように心がけること。reject する際は、技術的な誤りなどの正当な理由や、killer references を示すと共に、some words of encouragement を忘れないこと。
2. innovative idea を重視し、短期間で修正可能なミスを主原因として reject しないこと
3. 既存のアルゴリズムのバリエーションで、ちょっと思いついた、程度のものは、大規模な性能改善がない限りあまり重視しないこと
4. 実験に関しては、
 - i) 統計的に意味のある母数で実験がおこなわれていること
 - ii) 可能な範囲で標準的なデータセットとクエリを用いていること
 - iii) 更新とクエリシナリオがテストされていることが満足されていることを重視すること

私は約 250 の abstract に目を通して、14 本に willing の bid をたてました。当然ですが採録論文の投稿傾向と、投稿論文の投稿傾向は異なると思いますので、全投稿論文の abstract を参照できるのは、最新の研究動向を知る上で非常に有益だと思いました。ですので、メジャーな国際会議に継続的に採録されて存在感を示し、プログラム委員を委嘱されるようになって、世界の最新の研究動向に常に触れられる機会を持つようにすることが、研究を活発にまわしていくことにつながるのだと思いました。

大量の abstract を読んだ感想ですが、全体的に abstract の質は低い・低くなっているように思いました。abstract として、対象としている問題・課題、解決のアイデア、評価の骨子、などがちゃんと記載されていないものも少なくなかったです。少々極端ですが、abstract を読んでも title とほぼ同程度の情報しか得られないものもあり、title はかなり慎重につけているらしい、という印象を持つと同時に、abstract をもっとしっかり書くように気をつけなければいけない、と思いました。



ペーパーが割り当てられたのは、**2007.11.25** ですが、その日のうちに、**double-blind review violation** のペーパーが問答無用で **rejection** が決定しました。つまり、**double-blind review violation** のペーパーは11月末には **reject** が確定していたこととなります。それでも、最終通知は2月ですので、これはあまりにももったいないミスです。。。

ペーパーは、体裁が整ってしっかり書けているように見えるものもあれば、見ただけで、とても **SIGMOD** 採録レベルに達しているとは思えないものまで、予想に反して玉石混合という感じでした。

査読作業は、途中、**author feedback** の期間を経て、最初に **reject** のものが確定しました。基本的に、1ペーパーを2~3人のプログラム委員が査読していますが、全員が **reject** または **weak reject** で一致しているものは、**reject** になりました。なお、その際にも **Shasha** 教授からコメントが **helpful** であるかどうか、再度確認するように指示がありました。

次に、査読した全プログラム委員が、**strong accept/accept/weak accept** で一致しているペーパーが **Likely accept** になり、**SIGMOD** までによりよい論文にするために **author** がなすべきことが明確に示されているかどうか確認するように **Shasha** 教授から指示がありました。

最後に、上記のどちらでもないペーパーがプログラム委員間でのディベート対象になりました。ディベートに際して、

1. 査読した全プログラム委員が、**reject** 側、または **accept** 側になったら、議論を停止させること
2. 上記のような結論に至らなかった場合、その **paper** を **VLDB2008** の **rolover** として考慮すべきかどうか考えること。対象とする **paper** は、基本的なアイデアは良いが、完結させるには多くの作業が必要だと思われるものとする。

という指示がありました。

私が担当した論文に関する議論の中で参考にしたいことを以下にまとめます。

1. 査読したプログラム委員の **confidence** が **medium** と **low** で、採点も **weak accept** と **weak reject** が混ざっており、3人のプログラム委員が問題点と考えた見解(実験に使用したデータセットが小さすぎる)が一致したので、**VLDB** への **rolover** とした。
2. **author feedback** を無視したことが主な原因で、支持していたプログラム委員まで **reject** に回ってしまい、結果的に **reject** になったケースがありました。
→**author feedback** には真摯に対応した方が良いと思いました。
3. 2人 **reject** で、1人 **strong accept** の **paper** があり、最後まで誰も自分の **score** を変えずに議論が紛糾し、最終的には意見の一致は見ずに決裂しましたが、そのペーパーは結果的には **accept** になっていました。



自分の担当しているペーパーを **rollover** にした場合は、再度自分で査読する必要があります。**rollover** のペーパーは、**VLDB** に投稿する際には論文の他に、各プログラム委員の指摘事項に対する回答を添えて投稿しています。論文の条件付き採録の過程のような感じになっています。

以上ですが、まとめますと、プログラム委員・査読者として自分が感じたことは、最初にきちんと今回の査読指針が示されたこと、またそれに共感できたこと、が、今回の作業を進めるなかで非常に大きかったと思います。

また、今後も査読の機会があれば **Shasha** 教授のように、論文には積極的に良い点と見つけることを第一に、査読するよう心掛けたいと思いました。

また、この経験を通して、研究者としては、以下のことに今後は気をつけていきたいと思いました。

1. **paper** の内容を正しく簡潔に伝えられる **title** をつける
2. **abstract** には、この論文は何について書いてあるのか、どのような課題をどのように解決したのか、どのくらい効果があったのかなどをきちんと書く
3. **anonymity** を含めて、論文の体裁には細心の注意を払う
4. **author feedback** には真摯に対応する

なお、**SIGMOD2008** のプログラム委員をするにあたって、たくさんの方々にお世話になりました。この場をかりて、お礼申し上げます。ありがとうございました。



データ工学研究専門委員会からのお知らせ

今後の行事予定など

□ DE 研ホームページ(リニューアルしました): <http://www.ieice.org/iss/de/jpn/>

DE 研スタッフ連絡先: de-staff@mail.ieice.org

□ 研究会

6 月 DE 研究会 (PRMU 共催, [JDB フォーラム](#) と併設)

日時: 2008 年 6 月 19 日(木), 20 日(金)

場所: 小樽市民会館(北海道小樽市)

URL: <http://www.ieice.org/iss/de/jpn/jdb2008.html>

□ 国際会議

ACM SIGMOD/PODS

日時: 2008 年 6 月 9 日(月)~12 日(木)

場所: カナダ バンクーバー

URL: <http://www.sigmod08.org/>

□ VLDB

日時: 2008 年 8 月 24 日(月)~30 日(木)

場所: ニュージーランド オークランド

URL: <https://www.cs.auckland.ac.nz/research/conferences/vldb08>

□ DE 研の登録について

データ工学研究専門委員会(DE 研) の登録 (技術研究報告の予約購読) をよろしくお願ひします。

◎技術研究報告予約の案内ページ: <http://www.ieice.org/jpn/books/kenkyuuhoukoku.html>

◎登録 (予約購読) の申込書: <http://www.ieice.org/jpn/books/gihoumoushikomisho.html>

◎Web での申し込み: <http://www.ieice.org/jpn/books/gkenkyuform.html>