How to Write a Research Paper (or How to Graduate Quickly)?

DB Group Summer Seminar

Dongwon Lee
May 19, 2005

Announcement

- Over the summer, Prof. Mitra's and Prof. Lee's groups will have a joint DB seminar
- Goals:
  - Forum for practice talks
  - Learn what others are working on
  - Get fresh ideas from others' works
  - Find collaborators for your research
  - Get to know each other

Justification

- I am probably a qualified person to give a talk on this topic… because
  - I'm still STRUGGLING to publish
  - I do get rejections a lot :-(
  - I'm still learning from failures
- What's being presented here is purely my suggestion
- Take it or leave it – upto you !!

What is the Goal of a Research Paper?

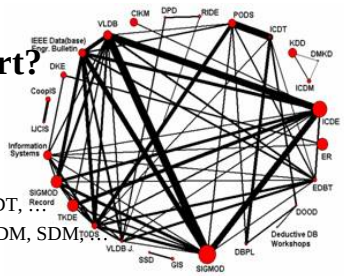
- Disseminate your ideas to others so that people appreciate/use/cite them
- **Graduate**… Of course
  - MS: need to write thesis to graduate…
  - Ph.D: "Publish or Perish"
- Without good publications…
  - No good job, no good career
  - And possibly no good life either
- GPA: nobody cares
  - Maintain about 3.0/4.0

Where to Start?

DB Conferences/Symposiums/Workshops (81)
ADB, ADBIS, ADBT, ADC, ARTDB, Berkeley Workshop, BNCOD, CDB, CIDR, CIKM, CISM, CISMOD, COMAD, COODBSE, CoopIS, DAISD, DANTE, DASFAA, DaWaK, DBPL, DBSEC, DDB, DDW, DEXA, DIWeB, DMDW, DMKD, DNIS, DOLAP, DOOD, DPDS, DS, EDBT, EDS, EFIS/EFDBS, ER, EWDW, FODO, FoIKS, FQAS, Future Databases, GIS, HPTS, IADT, ICDE, ICDM, ICDT, ICOD, IDA, IDC(W), IDEAL, IDEAS, IDS, IGIS, IWDM, IW-MMDBMS, JCDKB, KDD, KR, KRDB, LID, MDA/MDM, MFDBS, MLDM, MSS, NLDB, OODBS, OOIS, PAKDD, PKDD, PODS, RIDE, RIDS, RTDB, SBBD, SDM-SIAM, Semantics in Databases, SIGMOD, SSD, SSDBM, SWDB, TDB, TSDM, UIDIS, VDB, VLDB, WebDB, WIDM, WISE, XP, XSym
DB Journals (19)
ACM TODS, ACM TOIS, DKE, Data Base, DMKD, DPD, IEEE Data Eng. Bulletin, IEEE TKDE, Info. Processing and Management, Info. Processing Letters, Info. Sciences, Info. Systems, J. of Cooperative Info. Systems, J. of Database Management, JIIS, KAIS, SIGKDD Explorations, SIGMOD Record, VLDB J.

The list excludes Information Retrieval and Digital Library

Where to Start?



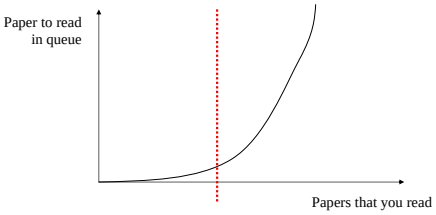
- Start from good ones:
  - DB: SIGMOD, VLDB, ICDE, EDBT, …
  - DB Theory: PODS, ICDT, …
  - Data Mining: KDD, ICDM, SDM, …
  - Modeling: ER, …
  - Information Retrieval: SIGIR, CIKM, …
  - Digital Library: JCDL, ECDL, CIKM, …
  - Web: WWW, WebDB, …
- Look at DBLP: http://www.informatik.uni-trier.de/~ley/db/

Where to Start?

- Don't be afraid to read **journal** papers
- DB field is a fast-moving discipline:
  - Latest techniques appear in conference/workshop
  - More mature work appears in journal
- Although longer than conference version, often easier to read
  - Lots of examples, figures, descriptions, …
- Examples:
  - ACM TODS, ACM TOIS, VLDB J., IEEE TKDE, ACM TOIT
  - ACM Computing Survey, C. ACM, SIGMOD Record, …

Reference Chase

- Don't trap into the "Exponential Reference Chase" problem



Symptoms

- After chasing relevant works that are increasing super-exponentially fast, you would feel…
- All relevant problems are ALREADY studied by someone else
  - Others have 1000+ history: Mathematics, Art, …
- Problem is too BROAD for me to tackle
  - Divide-n-conquer

How to Find the DARN Research Problem?

- Easy but non-helpful answer:
  - Read and think and read and think and…
- Subjective but MAYBE-helpful answer
  - **MAP** approach
  - **MATRIX** approach
  - **DELTA** approach
  - **DROP** approach

What I Call **M2D2**

1. MAP Approach

- To start a research, initially, you have to read a lot of papers anyway
- While reading those, why don't you analyze and summarize what you've read and put them into your own wording?
  - Good for a survey paper – a MAP for future readers
- To be publishable, your survey must have novel view-point, taxonomy, comprehensive analysis, or all of them
- Good target: ACM Comp. Survey, SIGMOD Record, ACM C.ACM, IEEE Computer, …

2. MATRIX Approach

- Now, You have read a lot of papers
- Draw a MATRIX on a specific problem, and map the paper that you read to cells of matrix
- At the end, non-filled cell is the missing work that no one has done
- But wait… first make sure that:
  - The hole is worthwhile to fill in
    - Doable (good as my dissertation topic?)
    - Value (what's good?)

Example: XML-Relational Conversion Problem

	Schema	Constraint	Query	View	Triggers	Security	Top-K	Temporal	Spatial
XML → Relational	O (40+)	O (5+)	O	O	O	O			
Relational → XML	O	O	O	O					

3. DELTA Approach

- Arguably easiest…
  - Pick one paper of your interest
  - Read a lot – more than 10 times
  - Find limitations and Extend it by DELTA
  - Prove or demonstrate that
    - The limitation that you pointed out is valid
    - Your suggestion improved the problem by DELA
- The more well-known work you choose, the harder to improve, but the better for your reputation…
  - Eg, "E.F. Codd's relational model is insufficient to handle semi-structured model because…"
- The bigger the DELTA is, the better your paper gets

Example: The optimal wedding problem

- When a person has a chance to date *K* persons, the optimal wedding algorithm is:
  - Date upto *K/3* persons
    - Let the best person among K/3 as *B* using a criteria *C*
  - Start dating again from *K/3+1* person, *p*
  - If *p* is better than *B* using *C*
    - Stop and Marry *p*
  - Otherwise, keep dating till *K-th* person
- How many ways can we improve this algo?

Possible DELTAs

- Parameters fitting:
  - How to determine *K*? Estimate?
  - How to determine C? Comparison?
- Scalability? K=10 vs. K=100,00? Sub-optimal?
- Question the assumptions:
  - Monogamy vs. Polygamy vs. N-gamy? (How to find n^{th} best spouse fast?)
  - Data distribution? Uniform/Poisson/Scale-free
- Application to another domain?
- System building?
- …

Which DELTA to Choose

- Pick the DELTA that is the most significant
- Some criteria are:
- Have practical values
  - Has motivational scenario as of NOW, or
  - Predicted to be useful in *N* years
- Non-trivial
- Hot topics:
  - Streaming, XML, Sensor, …

4. DROP Approach

(adopted from J. Widom's slides)

- Pick a simple but fundamental assumption underlying traditional database systems
  - **DROP it**
- Reconsider all aspects of data management and query processing
  - **Many Ph.D. theses**
  - **Prototype from scratch**

From http://www-db.stanford.edu/~widom/stream.ppt

Example: Two Stanford Projects

- The LORE Project
  - Dropped assumption: "*Data has a fixed schema declared in advance*"
  - Semi-structured data (→ XML)
- The STREAM Project
  - Dropped assumption: "*First load data, then index it, then run queries*"
  - Continuous data streams (+ continuous queries)

From http://www-db.stanford.edu/~widom/stream.ppt

Where to Submit?

- Top-down
  - Aim at the best conference in the field
  - If rejected, go to next-tier conference or symposium
  - If rejected, go to next…
- Bottom-up
  - Aim at workshop
  - If accepted, work more and aim at better one (symposium or 2nd-tier conference)
  - After making sure that the ideas mature enough, aim at the best conference

Avoid Some Notorious Venues

- "*Randomly generated paper got accepted to a conference… MIT Prank*" (slashdot, 2005)
  - http://pdos.csail.mit.edu/scigen/
  - Eg, The World Multi-Conference on Systemics, Cybernetics and Informatics (SCI)
- Along your career, you will get emails from unknown venues to submit a paper, to serve as PC, etc
  - Be careful if the venue is not well-known
  - Many of them are NON-REVIEWED, and Profit-Oriented event – no academic values what so ever !!

Facts on Paper Reviews

(adopted from J. Cho's slides)

- 3-4 reviewers per paper
- 10-20% acceptance rate for top-tier venues
  - Very competitive
- Criteria
  - Accept/Weak Accept
  - Neutral
  - Weak Reject/Reject
- One reject kills a paper
  - At least Accept, Weak Accept and Neutral

About Reviewers

- 15-20 papers per reviewer
- Reviewer cannot spend 5-10 hours per paper
  - 20 X 10 = 200 hours = (40 hours X 5) = 5 weeks!
  - No reviewers can afford this
- Give a good impression in 1-2 hours!
  - Impression matters the most
  - Content comes next!

WARNING: Of course, to start with, your main idea must be good to get into top-tier…

How to Give a Good Impression in 1-2 hours

1. Good introduction
   - Everyone reads it
   - If not interesting, people stop reading
2. Easy to read
   1. People should understand what you say
   2. Easy to confuse, difficult to understand
3. Build an excitement and a strong case
   1. What is good?
4. Broad reference
   1. Sometimes kills a paper
   2. Program committee members

Good Introduction

1. What's the problem?
2. Why is it important?
   - Mention some application, existing problems
3. Why is it difficult?
   1. Ask some not-very-obvious questions or explain naïve approach
4. What others did?
5. What's my contribution?
   1. Contribution bullet list (paper organization)
6. Build some excitement/surprise
   1. Keep reading! You will find something interesting later
7. Every word should be carefully picked

How to Write an Intro

1. Start with 5 bullets
   - What's the problem?
   - Why is it interesting?
   - …
2. 1-2 sentence answer to each question
3. Add more content
4. Spend enough time on intro
   1. Bullet points enough

Easy-to-Read Paper

- You can always make it complicate later

1. Lots of examples
2. Figures & Tables – Figure speaks !!
   - Summary of notations
3. Define models/architecture precisely
   - Explicitly write down assumptions
   - Input, output, property, goal function
4. Make a connection
   - Why this experiment?

Paper Organization (10 pages)

1. Introduction (2 pages)
2. Related Work (half page)
3. Framework (2 pages)
4. Main Ideas (3 pages)
5. Experiments (2 pages)
6. Conclusion (half page)
7. References (half page)

- Actual idea – only 3 pages!!!
  - Even tiny idea can turn into a good paper if you DEVELOP well

Importance of Personal Research Log

- Maintain personal research log
  - Sketch your research ideas into a writing
  - Update your ideas as time passes
  - Occasionally go back to old writings
- Prepare a short review for each paper that you read
  - Summary
  - Pros and cons
  - Limitations or problems
  - If needed, contact authors and ask questions
    - Usually authors are willing to discuss with their readers

Start Writing Early On…

- Even if you feel you are NOT ready yet
  - Your advisor will throw away your initial draft anyway
  - Your initial submission will be rejected anyway
- But you get
  - (good or bad) Experiences and learn from that
  - Writing sharpens your ideas and gives more ideas
  - Writing can be improved only via writing

Fabrication and Plagiarism

- *"Prominent Physicist Fired for Faking Data Research: Bell Labs says scientist 'recklessly' misrepresented work on microprocessors…" (2002, LA Times)*
  - http://www.latimes.com/news/science/la-sci-physicist26sep26.story
- *"Constantinos V. Papadopoulos got caught plagiarism at EUROPAR (1995)… 7 papers published and 8 under submission… all plagiarized from Technical Reports…"*
  - http://www.sics.se/europar95/plagiarism.html
- NEVER, EVER, do these – professional suicide !!

dbworld

- Be a member of dbworld newsgroup
  - http://www.cs.wisc.edu/dbworld/
  - Free membership
  - Keep track of DB-related news

References (available at)
http://nike.psu.edu/resources/advice/

- [2002] How to write a paper?, Junghoo Cho, UCLA
- [1996] David Dill's Advice on Choosing an Advisor (or) How to Survive as a Grad Student, David Dill
- [1996] How to Survive as a Graduate Student, Brian Noble, David Dill, Benli Pierce, Jay Sipelstein, Jonathan Shewchuck
- [1997] How to Choose a Thesis Advisor, Michael C. Loui
- [????] How to have your abstract rejected, Mary-Claire van Leunen and Richard Lipton
- [1994] Dissertation Advice, Olin Shivers
- [1999] Advice for Finishing that Damn Ph.D., Daniel M. Berry
- [1999] So long, and thanks for the Ph.D.!, Ronald T. Azuma
- [2001] How to Have a Bad Research Career, David A. Patterson