

Database/Data-Mining Research for Beginners*

Donghui Zhang
College of Computer & Information Science
Northeastern University
{donghui@ccs.neu.edu}

last modified: November 5, 2003

1 Introduction

Every field has its own conventions. For example, in the database community, we value the top conference papers (SIGMOD, etc.) more than journal papers. Thus a recent paper that I reviewed looked funny in the following sense: it has over 20 citations, none of which is a top database conference paper. The purpose of this document is to describe certain rules for graduate students who want to do database research.

2 How and Where to Submit

For every paper we write, we should follow a two-step procedure:

1. Send to a good database conference.
2. Once it's accepted, try to identify interesting problems to extend it (by at least 30%) and submit it to a journal.

There are hundreds of database conferences/symposiums/workshops. Although there are times that starting from small good workshops helps, we should try our best to submit to the most competitive conferences, as this increases the visibility to our work (more people attend/read these results).

Here's my suggested list, and we should try to submit only to them.

- **Best:** SIGMOD, VLDB
- **Very Good:** ICDE, EDBT, PODS
- **Good:** CIDR, ICDT, RIDE, WEBDB, WISE, etc.
- **Good:** (for spatial-temporal papers only) SSTD, GIS
- **Good:** (the topic is general, not only databases) SSDBM, CIKM.

*Disclaimer: the content of this document is the author's personal opinion. Follow the instructions at your own discretion.

For data mining, besides SIGMOD/VLDB, the list of best conferences should also include KDD. A good conference is ICDM.

The list of very good journals related to databases: TODS, VLDBJ, TKDE, Information Systems, etc. If you are not sure about a conference/journal you can always ask.

To get a journal version of a conference paper, in the cover page, cite the conference paper and say the extensions. Two common extensions are: (1) Cost models (especially for TODS) (2) Variations of the problem (e.g., approximate or continuous versions of the problem) or optimization of the techniques.

3 How to Find Paper, Author, Conference Information

Bookmark the DBLP page <http://www.informatik.uni-trier.de/ley/db/index.html>. Most database researcher visits that page several times a day.

Also, keep an eye on the DB Event page <http://nike.psu.edu/dbevents/>.

4 Some Routine Tasks

- Hang on your wall the schedule of conferences. Try to submit something to each conference. The schedule needs to be updated often.
- After the content of each major conference is known, scan through the paper list and abstract and get interesting papers.
- Visit the home page of selected researchers in your field often to see what they are doing.

5 How to Read a Paper Efficiently

- First, clearly identify the problem they address before reading their solutions.
- **STOP!** Try to solve the problem by yourself. You should at least find a straightforward solution.
- Read their solution and compare with yours, if any.
- **CONTINUE!** After reading the paper, besides modifying your maintained bib file and survey file (section 7), try to extend their solution to solve some other problem.

6 Tips for Presentation

You should have strong presentation skills whatsoever. Here are some tips in creating the talk slides and the presentation.

- Avoid small font in slides.
- Use figures and examples as much as possible.

- Avoid long sentences.
- Somewhere in the first couple of slides, clearly define the problem.
- Present it a couple of times to yourself first. Then present it to family/friends. The actual presentation is at the end.

7 Tips for Managing the Bibliography

You will read more and more papers. I used to keep all papers I had read in classified folders in a big cabinet. However, I found it is not a good method for two reasons. First, when I want to write a paper, I cannot find all the papers that I need to refer to. Second, for each paper that I need to refer to, I have to read it again to understand it and then write a small survey for it. Currently, I'm using the following technique:

- I keep the bibtex source for all papers I read in a file called *whole.bib*.
- I keep a survey file *survey.tex*, which briefly describes each paper in *whole.bib*. Of course, the survey file contains chapters and sections, etc.

Whenever I read a new paper which is interesting, I modify the bib file and the survey file.

To write the background section of a new paper, I more or less cut-and-paste from the survey file. For the new paper, we simply use *whole.bib* as the bib file. Although it contains many more entries than the current paper needs, Latex only extracts the items that are referred in the paper. If you really need to extract the referred items into a separate file, you can do it using the *extract.pl* script I wrote. This Perl script takes as input a filename. The file has one line: TAB separated citation keys. The script reads *whole.bib* file in the working directory, extracts all items with the given keys and prints them out on the screen. You should know how to redirect the output to a file.

You can gunzip, untar the attached file *reference.tgz*, which contains *whole.bib*, *survey.tex*, and *extract.pl*. Some related files are also included. Let me know if there is any question.

Some common errors related to the bib file:

1. No space between first name initial and last name. For instance, "D.Zhang". This will cause wrong ordering of papers cited. Should put a space there.
2. A single pair of brackets around *title*. This does not preserve the capitalization. For instance, "R-tree" will be generated as "r-tree". Should use double brackets.
3. No page number.
4. In a citation list in the previous text part, space between two citations. E.g. `\cite{BGO+96, AV88}`. In some cases, having a space between the two citations will cause error.