max planck institut
informatik

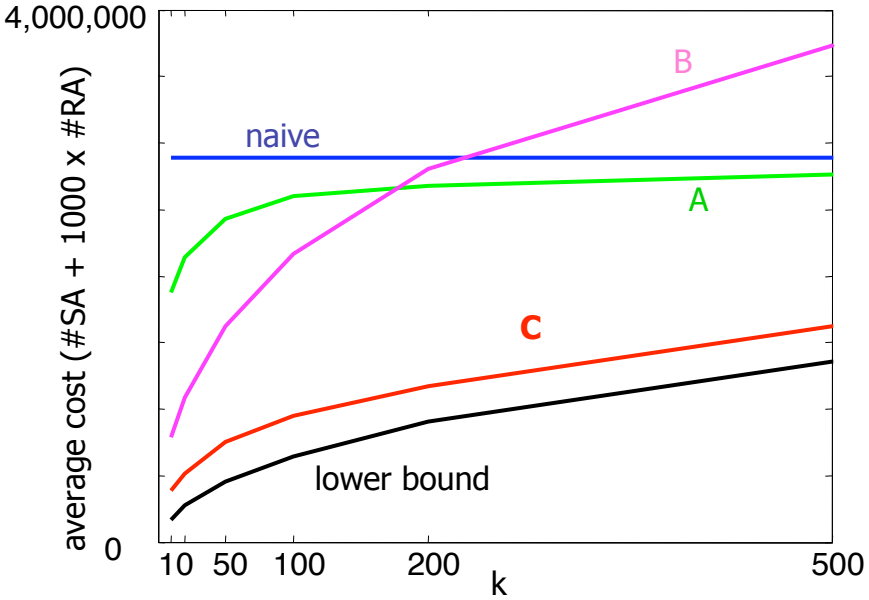
Archiving Experimental Results in Database Research

Ralf Schenkel

Databases and Information Systems Department
Semantic Search & Semistructured Data Group

Experiments in Computer Science?

Why is worst-case complexity not enough?



Algorithm A:
provably optimal complexity

Algorithm B:
provably optimal complexity

Naive solution:
sometimes better in practise

Algorithm C:
higher complexity, better on average

1. Constants do matter

2. Results need to be experimentally verified

In God we trust – and in charts

How to check experimental results?

„We implemented our algorithm in Java and tried three different data sets.“

[Then we threw away the results and made up the charts.]

Trust model involved (scientists don‘t cheat)

Mine is better than yours…

How to compare with existing results?

- ~~Kindly ask other authors for code~~
- Reimplement other (older) algorithms (often hardly documented, details missing)
- Find similar data

Improvements may be due to weak opponent code!

SIGMOD Experimental Repeatability

- Optional for 2008 conference
- Planned to get mandatory later
- Need to submit
  - Software to run the experiments
  - Data
  - Metadata (readme.txt)
- Committee checks if results in the paper match real results (10 people, 400 papers)
- Long-term goal: provide this for download

Providing Software

- Source or binary format?
  (avoid built-in solutions for benchmarks)
- Windows, Linux, MacOS, Solaris, S60, …
- Windows 95, 98, 2000, XP, 2003, Vista, …
- Java, C, C++, C#, Fortran, Modula-2, PHP, …
- Java 6, 5, 1.4, 1.1.8, JavaME, …
- Databases, compilers, toolkits, GUIs, …

Need to limit freedom!

Providing Data

- Old problems: standard benchmarks (Traveling salesman, text retrieval, …)
- New problems: do-it-yourself benchmarks (Searching in social networks, distributed retrieval in peer-to-peer-systems, …)
  - Is this a reasonable benchmark for the problem?
- Data format: plain text, database dump, XMLish, ISO-something…?

No solution yet

Providing Metadata

- ~~Ideally: „click on start to run.“~~

1. Buy a server with 64 CPUs.
2. Buy 32 gigabytes of main memory.
3. Buy 5 Terabytes external storage.
4. …

124. Run „doexp.sh | gnuplot | kpdf“

Need small test cases that run on off-the-shelf hardware

And then there are lawyers…

- Many big players among industry (Microsoft, IBM, Google, Yahoo!, …)
  - Not allowed to provide code („we implemented it in SQL Server“)
  - Often not allowed to provide data (remember the AOL query log disaster?)
  - But: some data only available for industry (2 years of Yahoo! query logs…)

Special rules for industry?
Penalty for other researchers?

Long-term archive: really, really quite long

- Who pays the bill for a long-term archive?
  - Authors keep things on their server
    (but people move, departments are closed…)
  - Global archive (SIGMOD, VLDB, ACM, IEEE, …)
    10TB/conference,
    200 conferences/year
    -> 2 PB/year
    and the journals (hit the publishers?)
- Shipping of data (want to download 1TB?)
- Maintenance, updates, errors, …

It‘s a long, long way to go

- Heavy debate among database people
- Everybody agrees: We need this
- Everybody agrees: We can‘t do this
- Outcome currently unclear