Computational Journalism: A Call to Arms to Database Researchers

Sarah Cohen  Public Policy, Duke U.
Chengkai Li  CSE, U. Texas Arlington
Jun Yang  CS, Duke U.
Cong Yu  Google Inc.

CIDR, January 2011
Crisis

- Traditional news media: fewer readers → lower ad revenue → fewer resources → less original investigative reporting
- Journalism’s watchdog function is in trouble

Quis custodiet ipsos custodes?
(Who will guard the guardians?)

- Who will hold governments, corporations, and powerful individual accountable to society?

http://www.dbgallery.co.uk/historys-whos-who/195869_socrates.html
Opportunity

- **Democratizing data:** more data are becoming publicly available

- Computation has a proven track record with big data

  - **Computational journalism**
    - Lower cost
    - Increase effectiveness
    - Broaden participation: democratizing data analysis
Fact checking

- Fact-checking is absurdly difficult, even if you know SQL and the databases are cleansed and documented

\[ U\text{-}check: \textit{a relational investigative tool for you} \]
- No knowledge of schema or SQL required

- But is this simply natural language querying (NLQ)?
Example: Giuliani’s adoption claim

- In the 2007 Republican presidential debate, Giuliani claimed that “adoptions went up 65 to 70 percent” in New York when he was in office.

Administration for Children’s Services was created in 1996

http://www.factcheck.org/elections-2008/levitating_numbers.html
Why U-check ≠ NLQ

- Claims often are vague and/or involve complex queries
- Users don’t expect one-click fact-checking with instant gratification
- Clarifying a claim and tweaking the way it presents data are instructive in their own right

An interactive interface that relies on user feedback
- Suggest possible SQL queries for user to choose
- To help user choose, show English translations, preview answers, ask questions...
Fact-check+

• Test how robust a claim is
  What’s the margin? Did it change over time?
  What if we compare with six instead of five districts?

• See if similar claims hold for different settings
  How does my district do in a similar comparison?
  How about median income instead of employment rate?

• Monitor a claim over time
  What if we revisit the comparison a year later?
  Can we get an alert when the streak is broken?

☞ Allow reuse of expertise/effort beyond a single story
Finding answers → finding questions

- U-check allows us to build up a “library” of datasets, queries leading to claims, and stories using them

☞ A Reporters’ Black Box
  - Learn “standard” query templates from the library and human experts
  - Run all templates on new/updated data to find claims that hold
  - Rank claims for further investigation by journalists

http://2.bp.blogspot.com/_5F-zDFdXIOY/SYe4qdS_GBI/AAAAAAAAAR4/BFQC7ioIPjE/s320/black-box.jpg
Vision: a cloud for the crowd

**Cloud**: aggregate/share computing resources
- Large-scale, real-time data analysis
  - E.g., map/reduce for machine translation, information extraction, reporters’ black box, etc.

**Crowd**: aggregate/share data, tools, and insights
- Leverage the crowd in simpler and more effective ways
  - An “optimizer” for the investigative process with crowdsourcing support
Example: crime-ridden LA City Hall?

Suppose many blogs seem to be talking about high crime rates around LA City Hall; what do you do?

- Verify information extraction results from blogs?
- Trace blogs back to sources: EveryBlock.com → LAPD public database
- Check individual crimes in zip code 90012
- LAPD’s geocoding software used 90012 as the default zip when a street address couldn’t be mapped!

威尔士和史密斯，“洛杉矶最高的犯罪率吗？不，只是洛杉矶警方的错误。”《洛杉矶时报》。2009年4月5日。
An investigative “optimizer”

- The investigative process is difficult to plan
- Can our system help plan it intelligently (incl. directing the crowd), in a goal-driven fashion, like a query optimizer?
  - Specify tasks declaratively
  - Identify mini-tasks that can be crowdsourced
  - Quantify cost-benefit of mini-tasks
  - Matching mini-tasks to users
  - Coordinate/reprioritize execution of mini-tasks
  - ...

Conclusion

- The need to save watchdog journalism is pressing
- You and I may hold the key
- Journalism is not only a consumer of technology, but it can also drive computer science
  - Our paper discusses more ideas and relevant research areas, but we have barely scratched the surface
- Don’t miss out working on something with a cause!