Dealing with Web Data: History and Look Ahead

Junghoo Cho
(UCLA)

Hector Garcia-Molina
(Stanford University)

Outline

- Digital Library Project
- Web crawling and our VLDB 2010 paper
- What has happened since then
- Open Challenges

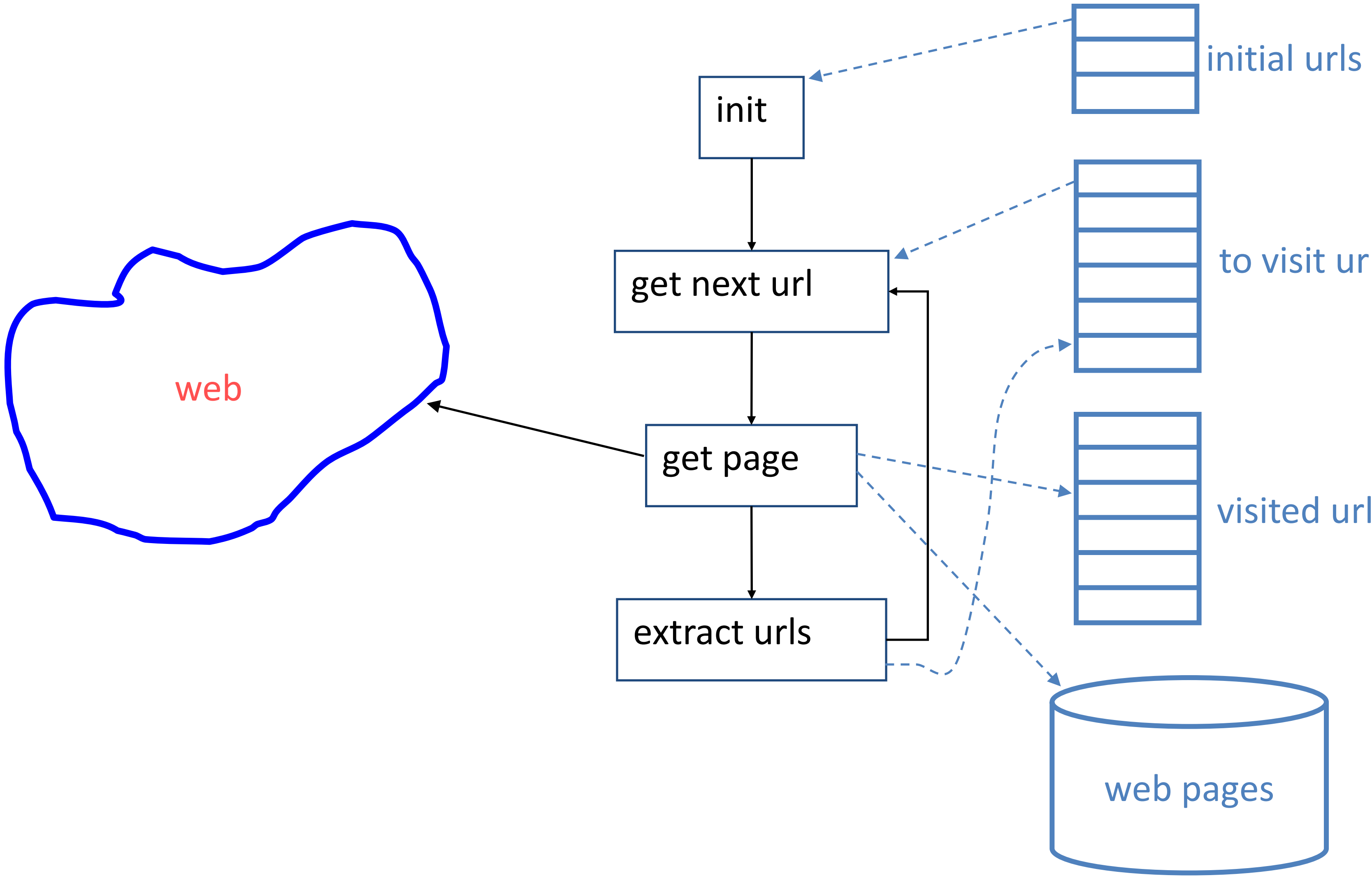
Digital Library Project

- NSF-funded research project, 1994-2004
- Develop technologies to integrate heterogeneous digital information for seamless, universal access
  - Data integration, clustering, and archival
  - Query and data translation
  - Data security, copyright protection
  - Mobile access

WebBase Project

- Collect, store, search and mine a significant portions of the Web
  - For Web search and mining research
  - Data repository for Web researchers
- WebBase crawler
- Backrub search engine, PageRank
  - Eventually became Google

What is a Crawler?



web

init

get next url

get page

extract urls

initial urls

to visit urls

visited urls

web pages

Crawling Issues

- Load at the site
  - Crawler should be unobtrusive to visited sites
- Load at the crawler
  - Download billions of Web pages in short time
- Page selection
  - Download “important” pages
- Page refresh
  - Refresh pages incrementally not in batch

VLDB 2000 Paper

- How to crawl the Web incrementally?
  - – Web evolution experiment
    - Active monitoring of half million Web pages
    - Poisson process as Web change model
  - – Incremental crawling policy and architecture
    - It is not always best to visit frequently changing pages more often
    - Crawler design choices and their impact

Since Our 2000 VLDB Paper

- Many follow-up work on Web crawling and Web evolution experiments
  - – 400 citations to our 2000 VLDB paper
- Web crawling
  - – 214 papers with keywords “Web crawler” in their title since 2000
    - Statistics are based on results from Google Scholar

Example of Recent Web Crawling Work

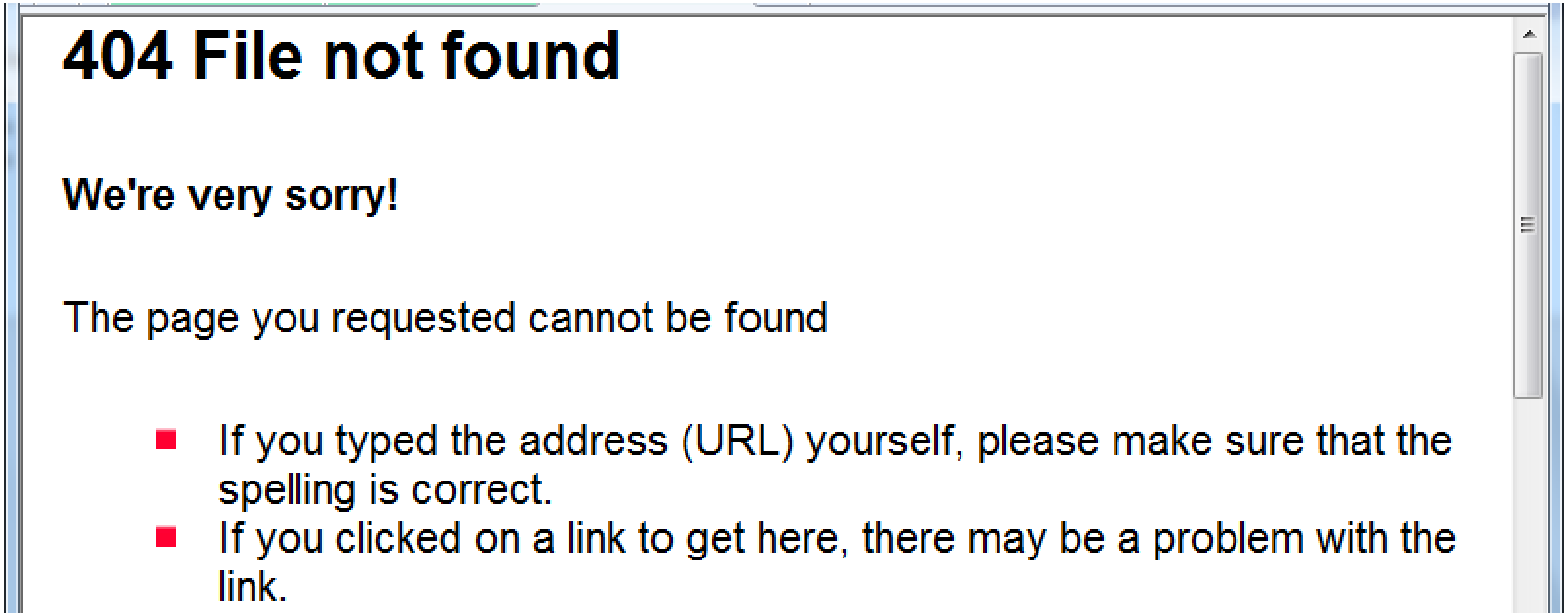
- "Crawl Ordering By Search Impact" by Pandey and Olston in WSDM 2008
  - Give high priority to the pages that are likely to handle the queries with few matching pages
- Google's sitemap protocol
  - Standard mechanism to inform Web crawlers of the URLs on the site and their modification date
  - Help crawlers discover pages to download and update
  - Based on our work on "Crawler-Friendly Web Servers" in 2000 PAWS

Follow-Up Work on Web Evolution

- 578 papers on Web evolution since 2000
  (According to Google Scholar)
- Example:
  - "On the Bursty Evolution of Blogspace" by Kumar et al. in WWW 2005
  - Study of exponential growth of graph connectivity within "blog" pages
  - Demonstrated formation of "micro-communities" within blogsphere and studied their time evolution

In Practice, This Resulted In … (1)

- No more “404 page not found” error
  - – 7% of search results are “broken” in 1999 [LG99]



404 File not found

We're very sorry!

The page you requested cannot be found

- If you typed the address (URL) yourself, please make sure that the spelling is correct.
- If you clicked on a link to get here, there may be a problem with the link.

In Practice, This Resulted In … (2)

- Significantly less indexing delay
  - Indexing delay of more than 6 months [LG99]
  - Important pages are indexed more than once a day by major search engines

In Practice, This Resulted In … (3)

- Significantly better coverage for popular queries
  - We get good results for most of navigational queries

But Things Are Not Done

- Complete paradigm shift in how Web is used
- Web as library vs Web as community
  - Twitter, Facebook, blogs, …
  - Exponential increase in generating and sharing personal and/or time-sensitive content
- We do not handle the "new" Web well
- New Challenges in
  - Scalability & performance issues
  - Helping users sift through data

Scalability & Performance (1)

- Ashton Kutcher at Twitter
  - 5.8M followers
  - 7 tweets per day on average

- Many other Twitter users like him
  - Barack Obama: 5.3M followers
  - Lady Gaga: 6.2M followers
  - Bill Gates: 1.5M followers
  - …

Scalability & Performance (2)

- Simple problem, but existing solutions are not adequate
  - Publish/subscribe system
  - Order of magnitude difference in data scale, distribution, and update
  - Twitter notorious for frequent outage
  - Problems are not unique to Twitter
  - Big companies develop their own in-house solutions
- Can we develop a general solution
  - Active ongoing research

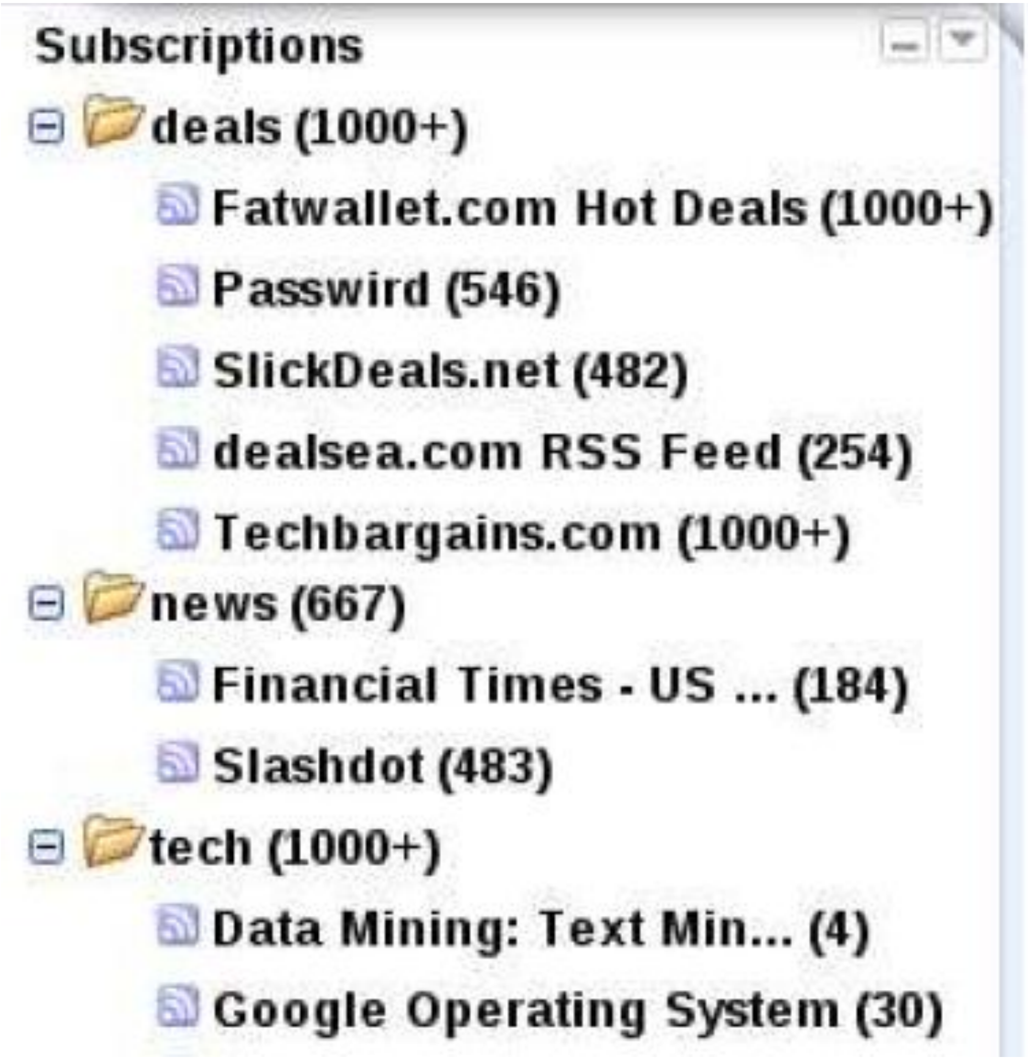
Thoughts on Review Process

- Excellent track record in evaluating scalability & performance work
- But some concerns
  - Preference to new and sophisticated ideas, not a new application of an old idea
    - "This has been done before by XXX"
    - "The solution is too simple"

New Challenges

- Scalability & performance issues
- Helping users sift through data

My Student’s Google Reader Page



Subscriptions

- deals (1000+)
  - Fatwallet.com Hot Deals (1000+)
  - Passwird (546)
  - SlickDeals.net (482)
  - dealsea.com RSS Feed (254)
  - Techbargains.com (1000+)
- news (667)
  - Financial Times - US ... (184)
  - Slashdot (483)
- tech (1000+)
  - Data Mining: Text Min... (4)
  - Google Operating System (30)

Existing Techniques are Limited

- Indicating sources to follow is not enough
  - Limited understanding of users and their interest
- Listing everything new is not enough
  - Limited understanding of information
- Simple keyword matching is not enough
  - Real-time search results are not satisfactory

How Humans Filter Information?

- My paper filtering process
  - Evaluate the source
    - What conference did it appear?
    - Who are the authors?
  - Evaluate the paper
    - Read title and abstract
  - Know myself
    - Is it the topic that I am interested in?

Replicating Human Filtering

- Can we replicate the human filtering process algorithmically?
- We need better models on
  - Users
  - Data
  - Sources
- PageRank is just a first-step to the solution

There Is Hope

- Richer meta data is available
  - Most information is tagged with its source
  - Most information is time-stamped
  - Information spread is traceable
- More data from diverse sources
  - Easier to learn general trend and pattern
  - It may be possible to ignore noise once the trend is learned
- Recent successes of probabilistic approaches
  - Probabilistic topic model as an example

Probabilistic Topic Model (1)

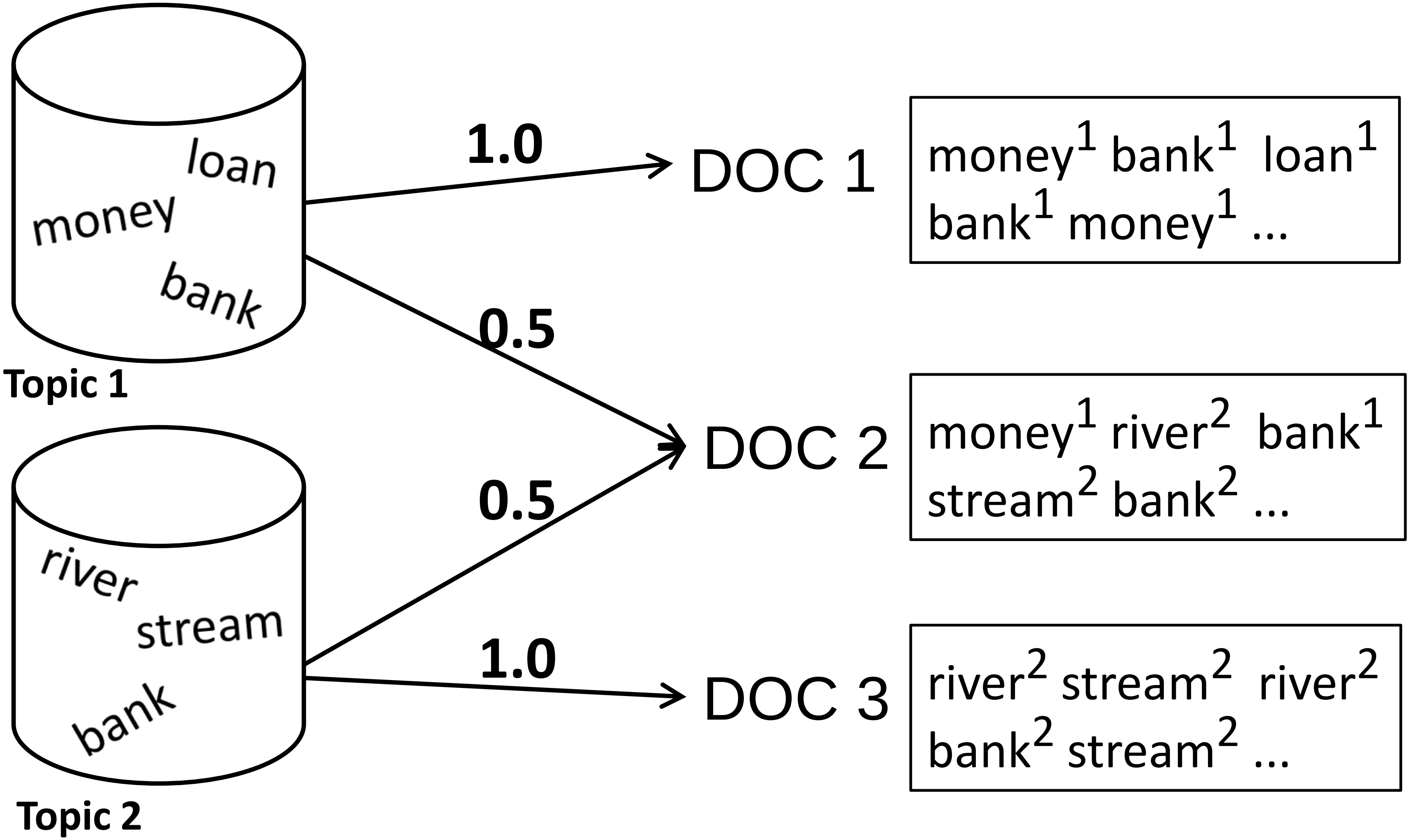
- Classify text into categories of topics
  - Decades-old problem with a large body of existing work, but with limited success
- Wide skepticism on papers on this topic until recently
  - "Yet another paper on document classification"
  - "Thousands of papers. Is there any more to study?"
  - "How much better can this be?"

Probabilistic Topic Model (2)

- In mid 2000, probabilistic latent semantic index (pLSI) and latent dirichlet analysis (LDA) were developed
  - The result blew away researchers in the field
- Model document generation as a probabilistic process
  - Infer the model parameters from available data

Probabilistic Document Model

$P(w|t)$ $P(t|d)$

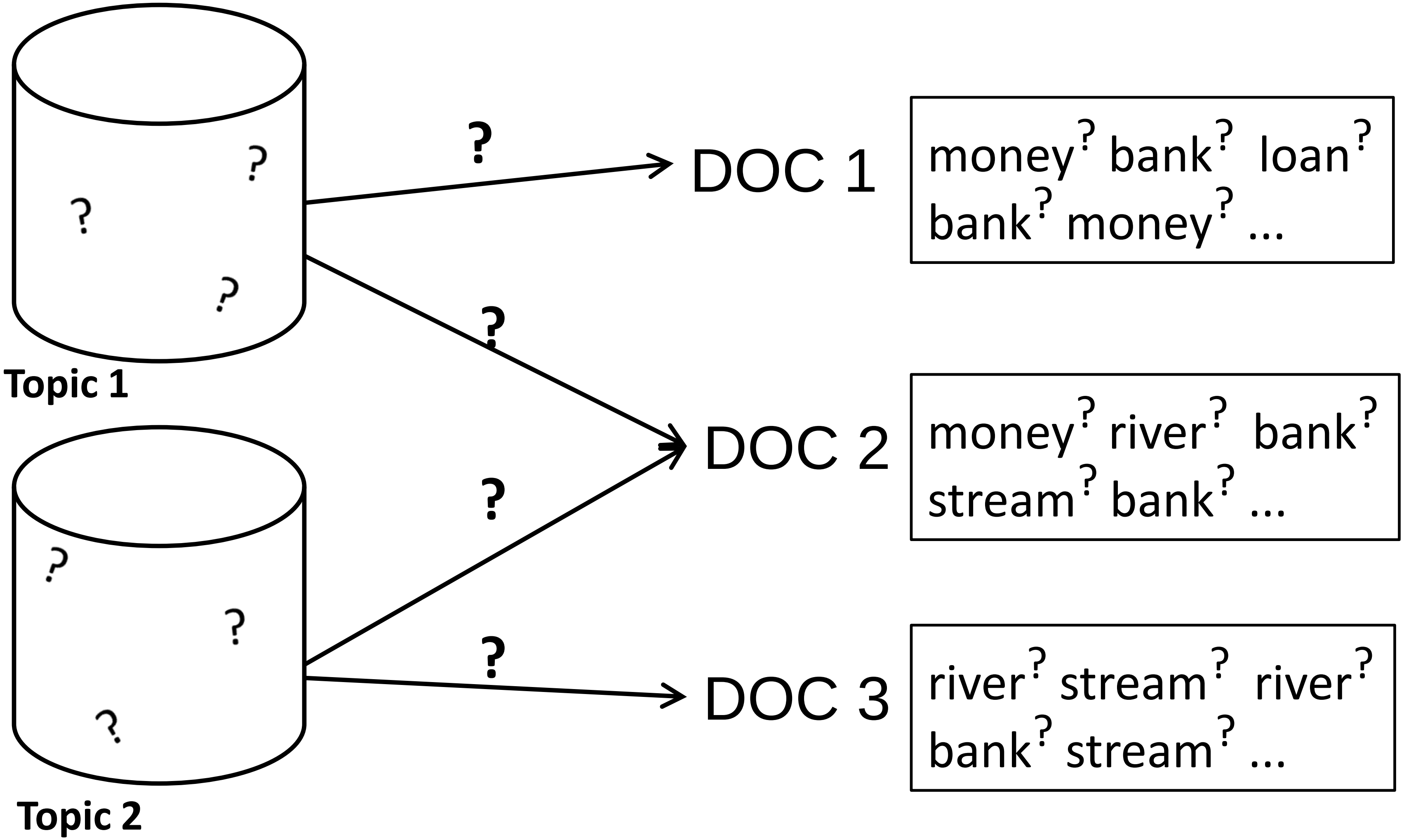


money^1 bank^1 loan^1 bank^1 money^1 ...

money^1 river^2 bank^1 stream^2 bank^2 ...

river^2 stream^2 river^2 bank^2 stream^2 ...

LDA as Topic Inference



money? bank? loan?
bank? money? ...

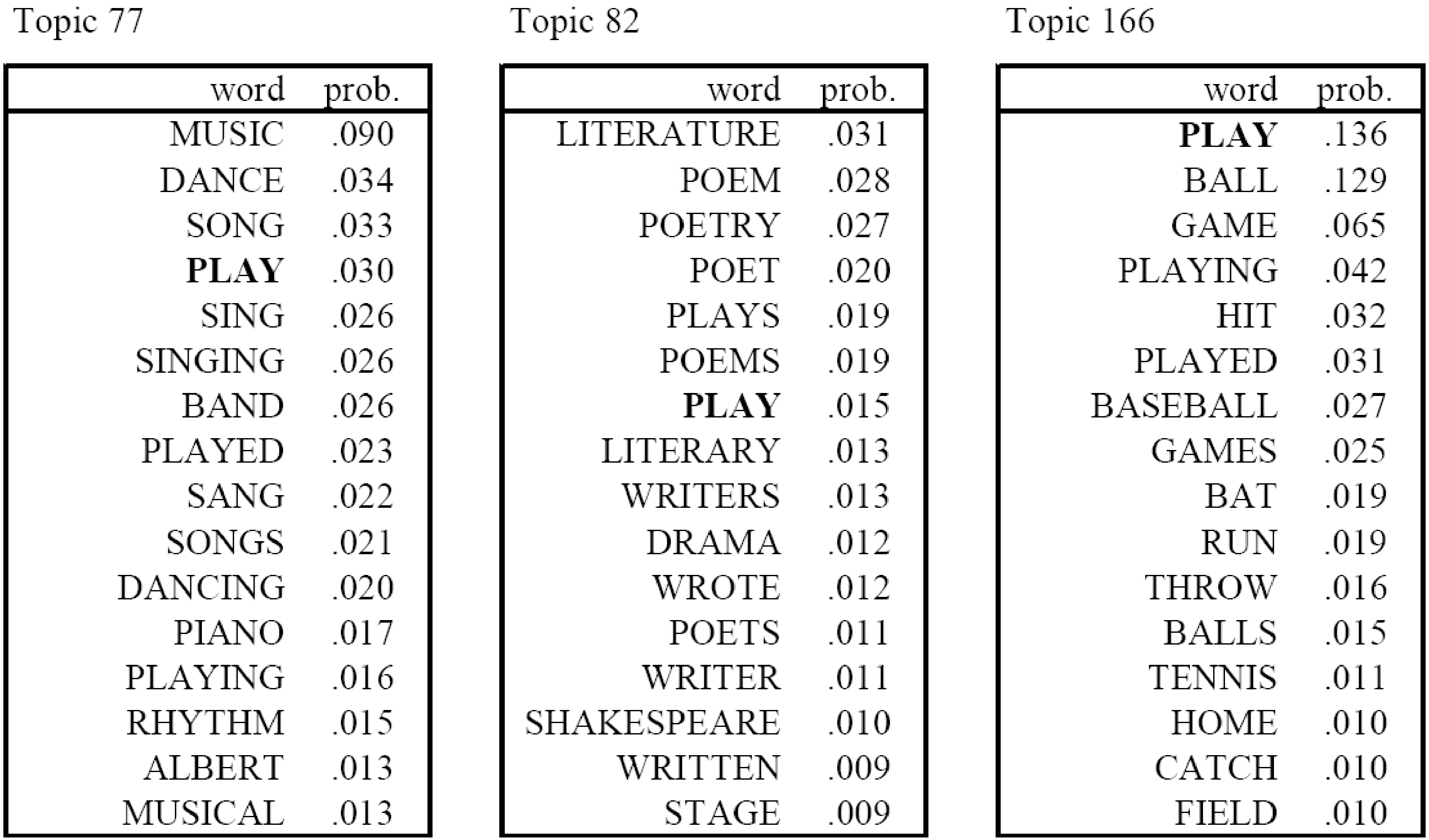
money? river? bank?
stream? bank? ...

river? stream? river?
bank? stream? ...

Results on Real Dataset [Steyvers 07]

- TASA corpus
  - 37,000 text passages from educational materials collected by Touchstone Applied Science Associates

Identified Topics



Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Unsupervised learning. Topics are learned without any training data.

Word Topic Assignment Result

- Document #29795
  … he was interested in another kind of music. He wanted to play^{077} the cornet. And he wanted to play^{077} Jazz …
- Document #1883
  … the actors must have the right playhouses and the playhouses must have the right audiences. We must remember that plays^{082} exist to be performed …

What Was Different?

- Strength of probabilistic approach
- Results are more “interpretable”
  - Resulting “numbers” are probabilities
- Resilient to input noise
  - Noise unavoidable for Web data
  - Outliers do not throw off the algorithm
- Apply probabilistic approach to other problems
  - Source modeling, user modeling, …

Thoughts on Our Review Process

- Terrible track record on papers on this topic
  - Where was the original PageRank paper published?
- Inherent challenge in working on this topic
  - Difficulty in providing quantifiable evidence
- What can we do to a better job?

Thank You

- We have done great work to build and support the constantly expanding Web and the users
- Many interesting challenges ahead
- Careful evaluation of our review process seem necessary
  - Support and encourage researchers who want to make an impact