Mining Deep Web Repositories

Nan Zhang, George Washington University
Gautam Das, University of Texas at Arlington and Qatar Computing Research Institute
<table>
<thead>
<tr>
<th></th>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>☀</td>
<td>Introduction</td>
</tr>
<tr>
<td>☀</td>
<td>Resource Discovery and Interface Understanding</td>
</tr>
<tr>
<td>☀</td>
<td>Technical Challenges for Data Mining</td>
</tr>
<tr>
<td>☀</td>
<td>Crawling</td>
</tr>
<tr>
<td>☀</td>
<td>Sampling</td>
</tr>
<tr>
<td>☀</td>
<td>Data Analytics</td>
</tr>
<tr>
<td>☀</td>
<td>Final Remarks</td>
</tr>
</tbody>
</table>
The Deep Web

Deep Web vs Surface Web

- Dynamic contents, unlinked pages, private web, contextual web, etc
- Estimated size: 91,850 vs 167 terabytes\(^1\), hundreds or thousands of times larger than the surface web\(^2\)

\(^1\) SIMS, UC Berkeley, How much information? 2003
Hidden Web Repositories

Diagram showing the flow of information from a Back-End DB through a Query Processing Module, to a Front-End Interface, and finally to the user. The interface includes fields for Make, Year, Model, Price, Mileage, Distance, and Listing Type, with specific options such as Ford, 2000 to 2006, F150, $1,000 to $10,000, and 20052.

Hidden Repository Owner
Web User
Deep Web Repository: Example I
Enterprise Search Engine’s Corpus

Unstructured data
Keyword search
Top-k

CDC - Asthma and Allergies - Prevention of Occupational ...
ASTHMA AND ALLERGIES. Prevention of Occupational Asthma: Introduction. ... Smith AM, Bernstein DI. Management of work-related asthma. ... www.cdc.gov/niosh/topics/asthma/OccAsthmaPrevention.html
[ More results from www.cdc.gov/niosh/topics/asthma ]

Lower Airway Rhinovirus Burden and the Seasonal Risk of Asthma Exacerbation,
Am J Respir Crit Care Med. 2011 Aug 4; [Epub ahead of print]
PMD: 21816938 [PubMed - as supplied by publisher]
Related citations

First Aid/CPR/AED - Professional Rescuers
... one- and two-rescuer); AED; Optional training in use of epinephrine auto-injectors and asthma inhalers available. Course ...

- Brand Name: Foradil Aerolizer
  Generic Name: Formoterol Fumarate Inhalation Powder
- Brand Name: Qvar
  Generic Name: Beclomethasone Dipropionate HFA
Exploration: Example I

Metasearch engine
- Discovers deep web repositories of a given topic
- Integrates query answers from multiple repositories
- For result re-organization, evaluate the quality of each repository through data analytics and mining
  - e.g., how large is the repository?
  - e.g., clustering of documents

Disease info
- CDC - Asthma and Allergies - Prevention of Occupational ...
  - www.cdc.gov/niosh/topics/asthma/OccAsthmaPrevention.html
  - Lower Airway Rhinovirus Burden and the Seasonal Risk of Asthma
  - PMID: 21816938 [PubMed - as supplied by publisher]

Treatment info
- Foradil Aerolizer: Formoterol Fumarate
- Qvar: Budesonide
- Asthma Treatment: Asthma Treatment Options
  - More-Get Info.
Example II
Yahoo! Auto, other online e-commerce websites

**Structured data**

**Form-like search**

**Top-1500**

---

### Vehicle

**Make**
- Select Make

**Model**
- Select Model

**Body Style**
- Any

**Year**
- Any to Any

**Price**
- Any to Any

**Mileage**
- Any to Any

---

![Yahoo! Autos](image)

<table>
<thead>
<tr>
<th>Picture</th>
<th>Year</th>
<th>Make and Model</th>
<th>Price</th>
<th>Mileage</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Car Image" /></td>
<td>2007</td>
<td><strong>BMW 335 xi</strong> Sedan, Black Sapphire Metallic, 3.0L I6, AUTO 6SPD, AWD, 4 door, Stock# 07130</td>
<td>$16,995</td>
<td>87,570 mi</td>
<td>Elizabeth, NJ Map</td>
</tr>
<tr>
<td><img src="image" alt="Car Image" /></td>
<td>2007</td>
<td><strong>BMW 335 Other Trim</strong> Convertible, Gold, Automatic</td>
<td>$17,600</td>
<td>13,500 mi</td>
<td><strong>Email Seller</strong></td>
</tr>
</tbody>
</table>

---

*Note: Images and data are placeholders for illustrative purposes.*
Exploration: Example II

Third-party analytics & mining of an individual repository
- Price distribution
- Price anomaly detection
- Classification: fake or real?

Third-party mining of multiple repositories
- Repository comparison
- Consumer behavior analysis

Main Tasks
- Resource discovery
- Data integration
- Single-/Cross-site mining

[Images of KAYAK, TERAPEAK, and pricegrabber logos]
Example III

Semi-structured data
Graph browsing
Local view

Picture from Jay Goldman, Facebook Cookbook, O'Reiley Media, 2008.
Exploration: Example III

For commercial advertisers:
- Market penetration of a social network
- "buzz words" tracking

For private detectors:
- Find pages related to an individual

For individual page owners:
- Understand the (relative) popularity or followers of one's own page
- Understand how new posts affect the popularity
- Understand how to promote the page

Main Tasks: Resource discovery and data integration less of a challenge, analytics and mining of very large amounts of data becomes the main challenge.
Summary of Main Tasks/Obstacles

- Find where the data are
  - Resource discovery: find URLs of deep web repositories
  - Required by: Metasearch engine, shopping website comparison, consumer behavior modeling, etc.

- Understand the web interface
  - Required by almost all applications.

- Mine the underlying data
  - Through crawling, sampling, and/or analytics
  - Required by: Metasearch engine, keep it real fake, price prediction, universal mobile interface, shopping website comparison, consumer behavior modeling, market penetration analysis, social page evaluation and optimization, etc.

Covered by many recent tutorials [Weikum and Theobald PODS 10, Chiticariu et al SIGMOD 10, Dong and Nauman VLDB 09, Franklin, Halevy and Maier VLDB 08]

Demoed by research prototypes and product systems

DBLife  WebTables  TextRunner

Arnetminer  Cazoodle  PatentsSearcher
Focus of This Tutorial

Brief Overview of:
- Resource discovery
- Interface understanding
- i.e., where to, and how to issue a search query to a deep web repository?

Our focus: Mining through crawling, sampling, analytics

Which individual search and/or browsing requests should a third-party explorer issue to the the web interface of a given deep web repository, in order to enable efficient data mining?
## Outline

- Introduction
- Resource Discovery and Interface Understanding
- Technical Challenges for Data Exploration
- Crawling
- Sampling
- Data Analytics
- Final Remarks
Resource Discovery

Objective: discover resources of “interest”
- Task 1: is an URL of interest?
  - Criteria A: is a deep web repository
  - Criteria B: belongs to a given topic
- Task 2: Find all interesting URLs

Task 1, Criteria A
- Transactional page search [LKV+06]
  - Pattern identification – e.g., “Enter keywords”, form identification
  - Synonym expansion – e.g., “Search” + “Go” + “Find it”

Task 1, Criteria B:
- Learn by example

Task 2
- Topic distillation based on a search engine
  - e.g., “used car search”, “car * search”
  - Alone not suffice for resource discovery [Cha99]
- Focused/Topical “Crawling”
  - Priority queue ordered by importance score
  - Leveraging locality
  - Often irrelevant pages could lead to relevant ones
    - Reinforcement learning, etc.

References:
Generally easy for keyword search interface, but can be extremely challenging for others (e.g., form-like search, graph-browsing).

What to understand?
- Structure of a web interface

Modeling language
- Flat model e.g., [KBG+01]
- Hierarchical model e.g., [ZHC04, DKY+09]

Input information
- HTML Tags e.g., [KBG+01]
- Visual layout of an interface e.g., [DKY+09]

[ZHC04] Z. Zhang, B. He, and K. C.-C. Chang, "Understanding Web Query Interfaces: Best-Effort Parsing with Hidden Syntax", SIGMOD 2004
Interface Understanding
Schema Matching

What to understand?
- Attributes corresponding to input/output controls on an interface

Modeling language
- Map schema of an interface to a mediated schema (with well understood attribute semantics)

Key Input Information
- Data/attribute correlation [SDH08, CHW+08]
- Human feedback [CVD+09]
- Auxiliary sources [CMH08]

Related Tutorials


# Outline

- Introduction
- Resource Discovery and Interface Understanding
- Technical Challenges for Data Mining
- Crawling
- Sampling
- Data Analytics
- Final Remarks
Assume that we are now given
- A URL for a deep web repository
- A wrapper for querying the repository (still limited by what queries are accepted by the repository – see next few slides)

What’s next?
- We still need to address the following challenge: which queries or browsing requests should we issue in order to efficiently support data mining?

Main source of challenge
- restrictions on query interfaces
- Orthogonal to the interface understanding challenge, and remains even after an interface is fully understood.
- e.g., how to estimate COUNT(*) through an SPJ interface
Problem Space and Solution Space

Traditional Heuristic Approaches
- e.g., seed-query based bootstrapping for crawling
- e.g., query sampling for repository sampling
- No guarantee on query cost, accuracy, etc.

Recent Approaches with Theoretical Guarantees
- e.g., performance-bounded crawlers
- e.g., unbiased samplers and aggregate estimators
- Techniques built upon sampling theory, etc.

Around 2000 ~ 2005 - now

Dimension 1: Task
- Analytics
- Sampling
- Crawling

Dimension 2: Interface
- Keyword Search
- Form-like Search
- Graph Browsing
Dimension 1. Task

- **Crawling**
  - Objective: download as many elements of interest (e.g., documents, tuples, metadata such as domain values) from the repository as possible.
  - Applications: building web archives, private directors, etc.

- **Sampling**
  - Draw sample elements from a repository according to a pre-determined distribution (e.g., uniform distribution for simple random sampling).
  - Why? Because crawling is often impractical for very large repositories because of practical limitations on the number of web accesses.
  - Collected sample can be later used for analytical processing, mining, etc.
  - Applications: Search-engine quality evaluation for meta-search-engines, price distribution, etc.

- **Data Analytics**
  - Directly support online analytics over the repository.
  - Key Task: efficiently answer aggregate queries (COUNT, SUM, MIN, MAX, etc.).
  - Overlap with sampling, but a key difference on the tradeoff of **versatility** vs. **efficiency**.
  - Applications: consumer behavior analysis, etc.
Why The Three Tasks?

Data mining can be enabled by

- Crawling: the crawled dataset can be treated as a local database
- Sampling: see the following slides for sample-based/facilitated data mining
- Data analytics: provides an API for data mining algorithms to call
Sample-Based / Facilitated Data Mining

Two general methods:

- Black-box approach: First generate a sample, and then apply data mining over the sample rather than the entire dataset.
  - Transparency can also be achieved at the OLAP level [LHY+08]
- White-box approach: use sample in selected steps (even preprocessing) of the data mining algorithm.

Surveys

Generic Methods

- **Input Reduction** (Black-box)
  - Sample from the input dataset the most important tuples for data mining

- **Divide-and-Conquer** (White-box)
  - Mine one sample set at a time
  - Combine results to produce the final mining results

- **Bootstrapping** (White-box)
  - Use sample to “guide” data mining over the entire dataset (e.g., as initialization settings)
Divide-and-Conquer: Windowing in ID3 [Qui86]
- first use a subset of the training set (i.e., a "window") to construct the decision tree
- then test it using the remainder of the training set, append mis-classified tuples to the window, and repeat the process until no mis-classification

Input Reduction: with stratified sampling [Cat91]
- esp. when the distribution of class labels is far from uniform in the training dataset
Sampling for Association Rule Mining

- **Bootstrapping**: find candidates from samples
  - first use samples to find approximate frequencies / candidate itemsets
  - then use the entire dataset to get the exact frequencies / verify candidates
  - possible to guarantee the discovery of all frequent itemsets (i.e., Las Vegas algorithm)
  - [AMS+96] [Toi96] [ZPL097] [LCK98] [CHH+05] [CGG10]
Sampling for Clustering

- **Bootstrapping**: use sample for initial settings
  - HAC on sample to bootstrap EM [MH98]

- **Input Reduction**
  - use sampling to neglect small clusters
  - density based sampling (oversample in sparse areas, undersample in dense ones) [PF00]
Keyword-based search
- Users specify one or a few keywords
- Common for both structured and unstructured data
- e.g., Google, Bing, Amazon.

Form-like search
- Users specify desired values for one or a few attributes
- Common for structured data
- e.g., Yahoo! Autos, AA.com, NSF Award Search.
- A similar interface: hierarchical browsing

Graph Browsing
- A user can observe certain edges and follow through them to access other users’ profiles.
- Common for online social networks
- e.g., Twitter, Facebook, etc.

A Combination of Multiple Interfaces
- e.g., Amazon (all three), eBay (all three).
Key Challenge
Restrictive Input Interface

- Restrictions on what queries can be issued
  - Keyword Search Interface: nothing but a set of keywords
  - Form-like Interface: only conjunctive search queries
    - e.g., List all Honda Accord cars with Price below $10,000
  - Graph Browsing Interface
    - only select one of the neighboring nodes

- We do not have complete access to the repository. No complete SQL support
  - e.g., we cannot issue “big picture” queries: e.g., SUM, MIN, MAX aggregate queries
  - e.g., we cannot issue “meta-data” queries: e.g., keyword such as DISTINCT (handy for domain discovery)
Key Challenge
Restrictive Output Interface

Restrictions on how many tuples will be returned
  - Top-k restriction leads to three types of queries:
    - **overflowing** (> k): top-k elements (documents, tuples) will be selected according to a (sometimes secret) scoring function and returned
    - **valid** (1..k element)
    - **underflowing** (0 element)
  - COUNT vs. ALERT
    - An alert of overflowing can always be obtained through a web interface
  - Page turn
    - Limited number of page turns allowed (e.g., 10-100 for Google)
      - Essentially the same as top-k restriction
  - Unlimited page turns
    - But a page turn also consumes a web access

A maximum of 3000 awards are displayed. If you did not find the information you are looking for, please refine your search.

Your search returned 41427 results. The allowed maximum number of results is 1000. Please narrow down your search criteria and try your search again.
Two ways to address the input/output restrictions

- Direct negotiation with the owner of the deep web repository
  - Crawling, sampling and analytics can all be supported (if necessary)
  - Used by many real-world systems - e.g., Kayak

- Bypass the interface restrictions
  - By issuing a carefully designed sequence of queries
  - e.g., for crawling: these queries should recall as many tuples as possible
    - or even “prove” that all tuples/documents returnable by the output interface are crawled.
  - e.g., for analytics: one should be able to infer from these queries an accurate estimation of an aggregate that cannot be directly issued because of the input interface restriction.
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
</tr>
<tr>
<td>2. Resource Discovery and Interface Understanding</td>
</tr>
<tr>
<td>3. Technical Challenges for Data Exploration</td>
</tr>
<tr>
<td>4. Crawling</td>
</tr>
<tr>
<td>5. Sampling</td>
</tr>
<tr>
<td>6. Data Analytics</td>
</tr>
<tr>
<td>7. Final Remarks</td>
</tr>
</tbody>
</table>
Overview of Crawling

- Motivation for crawling
  - Enable third-party web services - e.g., mash-up
  - A pre-processing step for answering queries not supported by the web interface
    - e.g., count the percentage of used cars which have GPS navigation; find all documents which contain the term “DBMS” and were last updated after Aug 1, 2011.
    - Note: these queries cannot be directly answered because of the interface restrictions.
  - Note the key differences with web crawling

- Taxonomy of crawling techniques
  - Interfaces: (a) (keyword and form-like) search interface, (b) browsing interface
  - Technical challenges: (1) find a finite set of queries that recall most if not all tuples (a challenge only for search interfaces), (2) find a small subset while maintaining a high recall, (3) issue the small subset in an efficient manner (i.e., system issues).

- Our discussion order
  - (a1), (a2), (b2), (*3)
Crawling Over Search Interfaces

(a1) Find A Finite Set of Search Queries with High Recall

- **Keyword search interface**
  - Use a pre-determined query pool: e.g., all English words/phrases
  - Bootstrapping technique: iterative probing [CMH08]

- **Form-like search interface**
  - If all attributes are represented by drop-down boxes or check buttons
    - Solution is trivial
  - If certain attributes are represented by text boxes
    - Prerequisite: attribute domain discovery
    - Nearly impossible to guarantee complete discovery [JZD11]
      - Reason: top-k restriction on output interface
      - \( k: \Omega(|V|^m); \) query cost: \( \Omega(m^2|V|^3) \)
      - Probabilistic guarantee achievable
    - Note: domain discovery also has other applications – e.g., as a preprocessor for sampling, or standalone interest.

---


Crawling Over Search Interfaces
(a2) How to Efficiently Crawl - Heuristics

- Motivation: Cartesian product of attribute domains often orders of magnitude larger than the repository size
  - e.g., cars.com: 5 inputs, 200 million combinations vs. 650,000 tuples
- How to use the minimum number of queries to achieve a significant coverage of underlying documents/tuples
  - Essentially a set cover problem (but inputs are not properly known before hand)
- Search query selection
  - Keyword search: a heuristic of maximizing #new_elements/cost [NZC05]
    - #new_elements: not crawled by previously issued queries
    - Cost may include keyword query cost + cost for downloading details of an element
  - Form-like search: find “binding” inputs [MKK+08]
    - Informative query template: grow with increasing dimensionality
    - Good news: #informative templates grows proportionally with the database size, not #input combinations.

Crawling Over Search Interfaces
(a2) How to Efficiently Crawl - Theoretical Bounds

Crawling Algorithms for Form-Like Search [SZT+12]
- $O(mn/k)$ for a numeric database.
- $U_1$ when there is only one categorical attribute.
- $n/k \sum_{i=1}^{m} \min(U_i, n/k) + \sum_{i=1}^{m} U_i$ for a categorical database.
- $U_1 + O(mn/k)$ for a mixed database with one categorical attribute.
- $n/k \sum_{i=1}^{m} \min(U_i, n/k) + \sum_{i=1}^{m} U_i + O((m\cdot\text{cat})\cdot n/k)$ for a mixed database with $\text{cat}$ (cat > 1) categorical attributes.

None of these can be improved beyond a constant factor!

---

Crawling Over Search Interfaces
(a2) How to Efficiently Crawl - Theoretical Bounds

Summary:

We pay $O(n/k)$ for each numeric attribute.

We pay $n/k \times \min(U_i, n/k) + U_i$ for each categorical attribute. Only exception: when there is only one categorical attribute.
Crawling Over Search Interfaces
(a2) How to Efficiently Crawl – Numeric DB

- **Baseline Algorithm:** Recursive equal-width binary split.
  - Problem: Worst-case query cost is domain-size dependent.
  - Pitfall: Many wasted queries - overflow and return no new tuple beyond previously seen.

- **Key Idea of Rank-Shrink:** Make sure no query is wasted.
  - each post-split query either is valid, or returns at least \( \frac{k}{4} \) new tuples.

- **How?** Consider the ordered set of \( k \) tuples returned by \( q \):
  - 1. The \( \lfloor k/2 \rfloor \)-th value shared by more than \( k/4 \) (identical) tuples?
  - 2. YES → ternary split, NO → binary split.
  - Worst-case query cost: \( \frac{12n}{k} \).
Crawling Over Search Interfaces

(a2) How to Efficiently Crawl - RANK-SHRINK for Numeric DB

Upper bound on query cost:
$20 \times m \times n / k$

Key Observation: No resolved query covers more than 1 tuple.

Implication: \( \exists \varepsilon = k/d > 0, \text{s.t. query cost} \geq dn/k \cdot \varepsilon. \)
Baseline strategy: Depth-first search
- Problem: Query cost (i.e., tree size) depends on domain size which can be unbounded.

Key Observation: “Essential” domain size is indeed bounded.
- 1 “Cover” Ai - i.e., query A_i = v_1, A_i = v_2, ..., A_i = v_{U_i} for all domain values for Ai respectively.
- 2 Ignore all v_j for which A_i = v_j returns valid.
- 3 At most min{U_i, n/k} values left in the essential domain.

Key Idea of Slide-Cover: Cover-Then-Slice
- 1 “Cover” to find the essential domains of all attributes.
- 2 “Slice” by performing DFS over a tree constructed from the essential domains.

Worst-case query cost
- U_1 when there is only one categorical attribute.
- n/k * Σ_{i=1}^{m} min(U_i, n/k) + Σ_{i=1}^{m} U_i otherwise.
Crawling Over Search Interfaces
(a2) How to Efficiently Crawl – Worst-Case Categorical DB

<table>
<thead>
<tr>
<th>A₁</th>
<th>A₂</th>
<th>⋯</th>
<th>Aₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>⋯</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>⋯</td>
<td>0</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋯</td>
<td>⋮</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>⋯</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>⋯</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>⋯</td>
<td>1</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋯</td>
<td>⋮</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>⋯</td>
<td>2</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋯</td>
<td>⋮</td>
</tr>
<tr>
<td>⋮</td>
<td>U − 1</td>
<td>⋯</td>
<td>U − 1</td>
</tr>
<tr>
<td>U − 1</td>
<td>0</td>
<td>⋯</td>
<td>U − 1</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋯</td>
<td>⋮</td>
</tr>
<tr>
<td>U − 1</td>
<td>U − 1</td>
<td>⋯</td>
<td>0</td>
</tr>
</tbody>
</table>
Crawling Over Browsing Interfaces
(b2) How to Efficiently Crawl

Technical problem
- Hierarchical browsing: Traverse vertices of a tree
- Graph browsing: Traverse vertices of a graph
  - Starting with a seed set of users (resp. URLs).
  - Recursively follows relationships (resp. hyperlinks) to others.
- Exhaustive crawling vs. Focused crawling

Findings
- Are real-world social networks indeed connected?
  - It depends – Flickr ~27%, LiveJournal ~95% [MMG+07]
- How to select “seed(s)” for crawling?
  - Selection does not matter much as long as the number of seeds is sufficiently large (e.g., > 100) [YLW10]

Using a cluster of machines for parallel crawling
  - Imperative for large-scale crawling
  - Extensively studied for web crawling
    - But are the challenges still the same for crawling deep web repositories?

Independent vs. Coordination
  - Overlap vs. (internal) communication overhead
  - How much coordination? Static vs. dynamic

Politeness, or server restriction detection
  - e.g., some repositories block an IP address if queries are issued too frequently – but how to identify the maximum unblocked speed?
<table>
<thead>
<tr>
<th>☑️</th>
<th>Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>☑️</td>
<td>Resource Discovery and Interface Understanding</td>
</tr>
<tr>
<td>☑️</td>
<td>Technical Challenges for Data Exploration</td>
</tr>
<tr>
<td>☑️</td>
<td>Crawling</td>
</tr>
<tr>
<td>☑️</td>
<td>Sampling</td>
</tr>
<tr>
<td>☑️</td>
<td>Data Analytics</td>
</tr>
<tr>
<td>☑️</td>
<td>Final Remarks</td>
</tr>
</tbody>
</table>
Overview of Sampling

Objective: Draw representative elements from a repository
  - Quality measure: sample skew
  - Efficiency measure: number of web accesses required

Motivating Applications
  - Unstructured data: use sample to estimate repository sizes [SZS+06], generate content summaries [IG02], estimate average document length [BB98, BG08], etc.
  - An interesting question: Google vs. Bing, whose repository is more comprehensive?
  - Structured data: rich literature of using sampling for approximate query processing (see tutorials [Das03, GG01])
  - An interesting question: What is the average price of all 2008 Toyota Prius@Yahoo! Autos?
  - Note (again): a sample can be later used for analytical purposes – e.g., data mining.

Central Theme
  - Skew reduction: make the sampling distribution as close to a target distribution as possible
    - Target distribution is often the uniform distribution – in this case, the objective is to make the probability of retrieving each document as uniform as possible.

References:
Sampling Over Keyword-Search Interfaces
Pool-Based Sampler: Basic Idea

- **Query-pool based sampler**
  - Assumption: there is a given (large) pool of queries which, once being issued through the web interface, can recall the vast majority of elements in the deep web repository
  - e.g., for unstructured data, a pool of English phrases

- **Two types of sampling process**
  - Heuristic: based on an observation that the query pool is too large to enumerate – so we have to (somehow) choose a small subset of queries (randomly or in a heuristic fashion) [IG02, SZS+06, BB98]
    - Problem: no guarantee on the “quality” (i.e., skew) of retrieved sample elements – e.g., if one randomly chooses a query and then randomly selects a document from the returned result [BB98], then longer documents will be favored over shorter ones.
  - Skew reduction: identify the source of skew and use skew-correction techniques, e.g., rejection sampling, to remove the skew.

- **Interesting observation: relationship b/w keyword and sampling a bipartite graph**

---


Sampling Over Keyword-Search Interfaces

Pool-Based Sampler: Reduce Skew

Doc1: This is the primary site for the Linux kernel source.

Doc2: Does Microsoft provide Windows Kernel source code for debugging purposes?


Doc4: The latest version of Windows OS Handbook is now on sale.

Sampling Over Keyword-Search Interfaces
Pool-Based Sampler: Reduce Skew

Doc1: This is the primary site for the Linux kernel source.

Doc2: Does Microsoft provide Windows Kernel source code for debugging purposes?


Doc4: The latest version of Windows OS Handbook is now on sale.

Sampling Over Keyword-Search Interfaces
Pool-Based Sampler: Remove Skew

Doc1: This is the primary site for the Linux kernel source.

Doc2: Does Microsoft provide Windows Kernel source code for debugging purposes.


Doc4: The latest version of Windows OS Handbook is now on sale.
Sampling Over Keyword-Search Interfaces
Pool-Based Sampler: Remove Skew

Doc1: This is the primary site for the Linux kernel.

Doc2: Does Microsoft provide Windows Kernel source code for debugging purposes?

Doc3: The latest version of the Windows Handbook is now on sale. It helps administrators become more effective.

Pool-free random walk [BG08]

- A graph model
  - Each element in the repository is a vertex
  - Two elements are connected if they are returned by the same query
- Random walk over the graph, two enabling factors:
  - Given an element, we can sample uniformly at random a query which returns the document. (YEA for almost all keyword search interfaces).
  - Given an element, we can find the number of queries which return the document (may incur significant query cost)
- Challenge 1: is the graph connected?
  - Note: the set of all possible queries which might return a document can be extremely large
    - $2^n$ queries for a document with $n$ words
  - Thus, we have to limit our attention to a subset of queries
    - e.g., only consecutive phrases
    - Problem: too restricted – disconnected graph, too relaxed – high cost for sampling
- Challenge 2: how to perform random walk?
  - Metropolis-Hastings algorithm

Sampling Over Form-Like Interfaces
Source of Skew

 Recall: Restrictions for Form-Like Interfaces
  o Input: conjunctive search queries only
  o Output: return top-k tuples only (with or without the COUNT of matching tuples)

 Good News
  o Defining “designated queries” no longer a challenge
  o e.g., consider all fully specified queries – each tuple is returned by one and only one of them
Bad News: A New Source of Skew

- We cannot really use fully specified queries because sampling would be really like search for a needle in a haystack
- So we must use shorter, broader queries
  - But such queries may be affected by the top-k output restriction
  - Skew may be introduced by the scoring function used to select top-k tuples
  - e.g., skew on average price when the top-k elements are the ones with the lowest prices

Basic idea for reducing/removing skew

- Find non-empty queries which are not affected by the scoring function – i.e., queries which return 1 to k elements
Sampling Over Form-Like Interfaces
COUNT-Based Skew Removal

\[ A_1 = 0 \& A_2 = 0 \]
\[ A_1 = 0 \& A_2 = 1 \]
\[ A_1 = 0 \& A_2 = 0 \& A_3 = 0 \]
\[ A_1 = 0 \& A_2 = 1 \& A_3 = 1 \]

Sampling Over Form-Like Interfaces
COUNT-Based Skew Removal

\[ \frac{3}{4} \times \frac{2}{3} \times \frac{1}{2} = \frac{1}{4} \]

[000] 001 010 011 100 101 110 111

Count=1
Count=2
Count=3

A1
A2
A3

Sampling Over Form-Like Interfaces
COUNT-Based Skew Removal

\[ \frac{3}{4} \times \frac{1}{3} = \frac{1}{4} \]

Sampling Over Form-Like Interfaces
Skew Reduction for Interfaces Sans COUNT

\[ \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \]

Sampling Over Form-Like Interfaces
Skew Reduction for Interfaces Sans COUNT

Solution: Reject with probability $1/2^h$, where $h$ is the difference with the maximum depth of a drill down

Note: Sampling is a challenge even when the entire graph topology is given
- Reason: Even the problem definition is tricky
  - What to sample? Vertices? Edges? Sub-graphs?

Methods for sampling vertices, edges, or sub-graphs
- Snowball sampling: a nonprobability sampling technique
- Random walk with random restart
- Forest Fire
- ...

What are the possible goals of sampling? [LF06]
- Criteria for a static snapshot
  - In-degree & out-degree distributions, distributions of weakly/strongly connected components (for directed graphs), distribution of singular values, clustering coefficient, etc.
- Criteria for temporal graph evolution
  - #edges vs. #nodes over time, effective diameter of the graph over time, largest connected component size over time,
Sampling Over Graph Browsing Interfaces
Unbiased Sampling

- Survey and Tutorials for random walks on graphs
  - [Lov93], [LF08], [Mag08]

- Simple random walk is inherently biased
  - Stationary distribution: each node $v$ has probability of $d(v)/(2|E|)$ of being selected, where $d(v)$ is the degree of $v$ and $|E|$ is the total number of edges – i.e., $p(v) \sim d(v)$

- Skew correction
  - Re-weighted random walk [VH08]
    - Rejection sampling
    - Or, if the objective is to use the samples to estimate an aggregate, then apply Hansen-Hurwitz estimator after a simple random walk.
  - Metropolis-Hastings random walk [MRR+53]
    - Transition probability from $u$ to its neighbor $v$: $\min(1, d(u)/d(v))/d(u)$
    - Stay at $u$ with the remaining probability
    - Leading to a uniform stationary distribution

---

[Mag08] M. Maggioni, Tutorial - Random Walks on Graphs Large-time Behavior and Applications to Analysis of Large Data Sets, MRA 2008.


<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Resource Discovery and Interface Understanding</td>
</tr>
<tr>
<td>Technical Challenges for Data Exploration</td>
</tr>
<tr>
<td>Crawling</td>
</tr>
<tr>
<td>Sampling</td>
</tr>
<tr>
<td>Data Analytics</td>
</tr>
<tr>
<td>Final Remarks</td>
</tr>
</tbody>
</table>
Objective: Directly estimate aggregates over a deep web repository

Motivating Applications

- Unstructured data: Google vs. Bing, whose repository is more comprehensive?
- Structured data: Total price of all cars listed at Yahoo! Autos?

Sampling vs. Data Analytics

- Data analytics requires the target aggregate to be known a priori. Samples can support multiple data analytics tasks.
- While samples may also be used to estimate (some, not all) aggregates, direct estimation is often more efficient because the estimation process can be tailored to the aggregate being estimated.

Performance Measures

- Quality measure: \( \text{MSE} = \text{Bias}^2 + \text{Var} \):
  - Reduction of both bias and variance.
- Efficiency measure: number of web accesses required
Analytics Over Keyword Search Interfaces
Leveraging Samples: Mark-and-Recapture

- Used for estimating population size in ecology.
- Recently used (in various forms) for estimating the corpus size of a search engine
  - Absolute size: [BFJ+06] [ZSZ+06] [LYM02]
  - Relative size (among search engines): [BB98] [BG08]

\[ \hat{m} = \frac{|C1| \times |C2|}{|C1 \cap C2|} \]

Note: only requires C1 and C2 to be uncorrelated - i.e., the fraction of documents in the corpus that appears in C1 should be the same as the fraction of documents in C2 that appear in C1

Problems with Mark-and-Recapture

**Problems**

- Correlation determination can be a tricky issue [BFJ+06]
  - e.g., C1: documents matching any five-digit number, C2: documents matching any medium frequency word – correlated
  - But – C1: documents matching exactly one five-digit number, C2 ... exactly one medium frequency word – little correlation

- Estimation bias
  - When using simple random samples, mark-and-recapture tends to be positively skewed [AMM05]

- (In-) Efficiency: at least an expected number of $m^{1/2}$ samples required for a population of size $m$

Analytics Over Keyword Search Interfaces
An Unbiased Estimator for COUNT and SUM

Objective: perform analytics over a search engine’s user query log, based on the auto-completion feature provide by the search engine (essentially an interface with prefix-query input restriction and top-k output restriction)

When random walk stops at node $x$

$$E \left[ \frac{1}{p(x)} \right] = \sum_{x \text{ is marked}} p(x) \cdot \frac{1}{p(x)} = \# \text{ of marked nodes}$$

Analytics Over Form-Like Interfaces
An Unbiased Estimator for COUNT and SUM

Basic Ideas
✓ Continue drill down till valid or underflow is reached
✓ Size estimation as $\frac{|q|}{p(q)}$ (Hansen-Hurwitz Estimator)
✓ **Unbiasedness** of estimator

$$E \left[ \frac{|q|}{p(q)} \right] = \sum_{q \in \Omega_{TV}} p(q) \cdot \frac{|q|}{p(q)} = m$$
Analytics Over Form-Like Interfaces
An Unbiased Estimator for COUNT and SUM


Basic Ideas
✓ Continue drill down till valid or underflow is reached
✓ Size estimation as $\frac{|q|}{p(q)}$ (Hansen-Hurwitz Estimator)
✓ Unbiasedness of estimator $E\left[\frac{|q|}{p(q)}\right] = \sum_{q \in \Omega_{TV}} p(q) \cdot \frac{|q|}{p(q)} = m$
Weight Adjustment
- Addresses low-level low-cardinality nodes

Divide-and-Conquer
- Addresses deep-level dense nodes

Stratified Sampling [LWA10]

Adaptive sampling
- e.g., adaptive neighborhood sampling: start with a simple random sample, then expand it with adding tuples from the neighborhood of sample tuples [WA11]

Analytics Support for Data Mining Tasks
- Frequent itemset mining [LWA10, LA11], differential rule mining [LWA10]

Observation: uniqueness of analytics over graph browsing
- Aggregates over a graph browsing interface may be defined on not only the underlying tuples (i.e., each user’s information), but also the graph topology itself (i.e., relationship between users)
- Examples: Graph cut, size of max clique, other topological measures

Implication of the uniqueness
- It is no longer straightforward how a sample of nodes can be used to answer aggregates
- Efficiency and accuracy of analytics now greatly depend on what topological information the interface reveals, e.g.,
  - Level 1: a query is needed to determine whether user A befriends B.
  - Level 2: a query reveals the list of user A’s friends.
  - Level 3: a query reveals the list of user A’s friends, as well as the degree of each friend.
Analytics Over Graph Browsing Interfaces
Relationship with Graph Testing

 PREFIX Graph Testing [GGR98, TSL10]
 o Input: a list of vertices
 o Interface: a query is needed to determine if there is an edge between two vertices
 o Objective: Approximately answer certain graph aggregates (e.g., k-colorability, size of max clique) while minimizing the number of queries issued.

 PREFIX Differences with Graph Testing
 o The list of vertices is not pre-known
 o More diverse interface models
 o More diverse aggregates
   • e.g., on user attributes
   • e.g., defined over a local neighborhood

Example: k-colorability [GGR98].
A simple algorithm of sampling $O(k^2 \log(k/\delta)/\varepsilon^3)$ vertices and testing each pair of them can construct a k-coloring of all n vertices such as at most $\varepsilon n^2$ edges violate coloring rule.

Outline

- Introduction
- Resource Discovery and Interface Understanding
- Technical Challenges for Data Exploration
- Crawling
- Sampling
- Data Analytics
- Final Remarks
Conclusions

- **Challenges**
  - Resource discovery
  - Interface understanding
  - Data exploration

- **Enabling Data Mining**
  - Tasks: Crawling, Sampling, Analytics
  - Interfaces: Keyword search, form-like search, graph browsing

**Traditional Heuristic Approaches**
- e.g., seed-query based bootstrapping for crawling
- e.g., query sampling for repository sampling
- No guarantee on query cost, accuracy, etc.

**Recent Approaches with Theoretical Guarantees**
- e.g., performance-bounded crawlers
- e.g., unbiased samplers and aggregate estimators
- Techniques built upon sampling theory, etc.
Open Challenges

- Is the black-box approach still viable?
  - high cost of acquiring samples => significantly smaller sample size
  - poor performance of small-sized simple random sample [PK99]

- Two key challenges
  - Deeper integration of sampling and data mining algorithms
  - Workload-aware sampling / aggregate estimation algorithms for deep web databases
Open Challenges

- **Website-Imposed Challenge**
  - Dynamic data - when aggregates change rapidly
    - e.g., Twitter, financial data, etc.
  - Hybrid of interfaces
  - Many others...

- **Privacy Challenge**
  - From an owner’s perspective: should aggregates be disclosed?
  - This challenge forms a sharp contrast with most existing work on data privacy (which focuses on **protecting** individual tuples while properly **disclosing** aggregate information for analytical purposes)
  - Here we must **disclose** individual tuples while **suppressing** access to aggregates
  - Recent work: dummy tuple insertion [DZDC09], correlation detection [WAA10], randomized generalization [JMZD11], adaptive query processing [ZZD12]

References


[CHH+05] S Cong, J Han, J Joeflinger, D Padua, A sampling-based framework for parallel data mining, PPoPP 2005.


References

- [LHY+08] X Li, J Han, Z Yin, J-G Lee, Y Sun, "Sampling cube: a framework for statistical OLAP over sampling data", SIGMOD 2008.


[YHZ+10] X. Yan, B. He, F. Zhu, J. Han, "Top-K Aggregation Queries Over Large Networks", ICDE, 2010


Thank you

Questions?

Contact: nzhang10@gwu.edu, gdas@uta.edu