Optimizing Information Extraction over Evolving Text

Fei Chen
Database Group
Computer Science Department
University of Wisconsin – Madison
Date: Wed Dec 20 08:57:00 EST 2000
From: Kay Mann <kay.mann@enron.com>
To: Suzanne Adams <suzanne.adams@enron.com>
Subject: Re: GE Conference

Did Sheila want Scott to participate?

80% of the world’s data, and growing!
Exploiting Text by Information Extraction (IE)

**Disease News**

Jan 31: Turkey confirmed an incident of foot&mouth disease.

Jan 30: H1N1 identified in California Turkey Flock.

**SELECT location**

**FROM Incidents**

**WHERE disease = ‘H1N1’**

**IE program**

<table>
<thead>
<tr>
<th>disease</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>foot&amp;mouth</td>
<td>Turkey</td>
</tr>
<tr>
<td>H1N1</td>
<td>California</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Data Mining**

**Web Search**

**Advertising**

**Monitoring**

**Visualizing**

Countries: US, Mexico, China, Canada.
Many players

- AI/DB/DM/IR/NLP/Web communities: CMU, Columbia, Cornell, IIT, JHU, Max-Panck, PSU, Stanford, UCLA, UIUC, UMass, UMichigan, USC, UT Austin, UToronto, UWashington, UWisconsin...

- industry: AT&T, Google, HP Labs, IBM, Microsoft, Yahoo!...

Mainly centered on improving extraction accuracy

Recent work starts to consider improving runtime...
# Current Work on Improving Runtime

## Pruning documents
- [Agichtein et al. ICDE-03](#)
- [Etzioni et al. WWW-04](#)
- [Ipeirotis et al. SIGMOD-06](#)

## Efficient pattern match
- [Gravano et al. VLDB-01](#)
- [Cho et al. ICDE-02](#)
- [Chandel et al. ICDE-06](#)

## Parallel processing
- [Gruhl et al. IBMJSJ-04](#)
- [Lin SIGIR-09](#)

## Relational-style optimization
- [Shen et al. VLDB-07](#)
- [Reiss et al. ICDE-08](#)

## IE over evolving text
- [Chen et al. ICDE-08](#)
- [Chen et al. SIGMOD-09](#)
- [Chen et al. TechReport-10](#)
IE over Evolving Text: An Example

Disease News
Jan 31: Turkey confirmed an incident of foot & mouth disease.
Jan 30: H1N1 identified in California Turkey Flock.

Alert me if there is an H1N1 incident

Incidents

<table>
<thead>
<tr>
<th>disease</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>foot &amp; mouth</td>
<td>Turkey</td>
</tr>
<tr>
<td>H1N1</td>
<td>California</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Recent Incidents

Flu Situation Update
Influenza activity remained at the same levels.

Reduction of Inventory at the

Disease News
Feb 1: A new H5N1 case confirmed in Indonesia.
Jan 31: Turkey confirmed an incident of foot & mouth disease.
Jan 30: H1N1 identified in California Turkey Flock.

diseases

<table>
<thead>
<tr>
<th>disease</th>
<th>location</th>
</tr>
</thead>
<tbody>
<tr>
<td>foot &amp; mouth</td>
<td>Turkey</td>
</tr>
<tr>
<td>H1N1</td>
<td>California</td>
</tr>
<tr>
<td>H5N1</td>
<td>Indonesia</td>
</tr>
<tr>
<td>foot &amp; mouth</td>
<td>Turkey</td>
</tr>
<tr>
<td>H1N1</td>
<td>California</td>
</tr>
</tbody>
</table>

...
IE over Evolving Text: Another Example

[Image of a webpage showing the DBLife portal for the database community]
<table>
<thead>
<tr>
<th>Researcher 1</th>
<th>Researcher 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joseph M. Hellerstein</td>
<td>Minos N. Garofalakis</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Researcher 1</th>
<th>Researcher 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joseph M. Hellerstein</td>
<td>Ion Stoica</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Researcher 1</th>
<th>Researcher 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joseph M. Hellerstein</td>
<td>Rajeev Rastogi</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Many Other Applications of IE over Evolving Text

- **Impliance @ IBM**
  - manages information on enterprise intranets
  - finds the latest information

- **IWP @ Univ. of Washington and YAGO @ MPI**
  - extract structures from Wikipedia
  - keep extracted structures up-to-date

- **Blogscope @ Univ. of Toronto**
  - monitors the blogosphere

- ...
No Good  Current Solution

Researchers:
- Joseph M. Hellerstein
- Minos N. Garofalakis
- Ion Stoica
- Rajeev Rastogi

Conference Pages:
- day 1
  - IE
  - Joseph M. Hellerstein
  - Minos N. Garofalakis
- day 2
  - IE
  - Joseph M. Hellerstein
  - Ion Stoica
- day 3
  - IE
  - Joseph M. Hellerstein
  - Rajeev Rastogi

Group Pages:
- DBworld mailing list
- DBLP
- Google Scholar
No Good Current Solution

- **Inefficient**
  - e.g., takes 8+ hours in DBLife everyday

- **Unsuitable for time-sensitive applications**
  - e.g., stock and auction
  - cannot finish extracting in time

- **Unsuitable for interactive debugging**
  - long debug loop
My Dissertation Contributions

Developed efficient IE solutions over evolving text that
- recycle previous IE efforts
- guarantee correctness
- deal with large text corpora
- deal with large IE programs
- deal with complex learning-based IE programs

Results in [ICDE-08], [SIGMOD-09] and [TechReportA-10]
Can we get $M_4$ faster than applying $E$ to $S_4$ from scratch?
Cyclex [ICDE08]: Recycling Extraction

slicdeals.net (day 1)

Hot Deals
Dec 10: Sony TV $850 at Amazon.
Dec 9: iPhone $199 at BestBuy.

Deals

<table>
<thead>
<tr>
<th>product</th>
<th>seller</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>Amazon</td>
</tr>
<tr>
<td>iPhone</td>
<td>BestBuy</td>
</tr>
</tbody>
</table>

slicdeals.net (day 2)

Hot Deals
Dec 11: All Dell LCDs $299.
Dec 10: Sony TV $850 at Amazon.
Dec 9: iPhone $199 at BestBuy.
Expired!

Deals

<table>
<thead>
<tr>
<th>product</th>
<th>seller</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCDs</td>
<td>Dell</td>
</tr>
<tr>
<td>TV</td>
<td>Amazon</td>
</tr>
<tr>
<td>iPhone</td>
<td>BestBuy</td>
</tr>
</tbody>
</table>
Not Always Correct to Copy Everything!

E extracts a deal if (a) product & seller are within a window of at most 20 chars (b) no “Expired!” is within 10 chars around the window.

slickdeals.net (day 1)

<table>
<thead>
<tr>
<th>Hot Deals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 10: Sony TV $850 at Amazon.</td>
</tr>
<tr>
<td>Dec 9: iPhone $199 at BestBuy.</td>
</tr>
</tbody>
</table>

slickdeals.net (day 2)

<table>
<thead>
<tr>
<th>Hot Deals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 11: All Dell LCDs $299.</td>
</tr>
<tr>
<td>Dec 10: Sony TV $850 at Amazon.</td>
</tr>
<tr>
<td>Dec 9: iPhone $199 at BestBuy.</td>
</tr>
<tr>
<td>Expired!</td>
</tr>
</tbody>
</table>

Wrong!
Extractor Properties: Context

- Many extractors only examine small "context windows" surrounding a mention to extract the mention.

E extracts a deal if:
(a) product & seller are within a window of at most 20 chars
(b) no "Expired!" is within 10 chars around the window.

Hot Deals
Dec 11: All Dell LCDs $299.
Dec 10: Sony TV $850 at Amazon.
Dec 9: iPhone $199 at BestBuy.
Expired!

- The text outside the context of a mention m is irrelevant for E to extract m.
Revisit The Example

E extracts a deal if (a) product & seller are within a window of at most 20 chars 
(b) no “Expired!” is within 10 chars around the window.

d slickdeals.net (day 1)

<table>
<thead>
<tr>
<th>Hot Deals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 10: Sony TV $850 at Amazon.</td>
</tr>
<tr>
<td>Dec 9: iPhone $199 at BestBuy.</td>
</tr>
</tbody>
</table>

E’s context = 10 chars

p

d slickdeals.net (day 2)

<table>
<thead>
<tr>
<th>Hot Deals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 11: All Dell LCDs $299.</td>
</tr>
<tr>
<td>Dec 10: Sony TV $850 at Amazon.</td>
</tr>
<tr>
<td>Dec 9: iPhone $199 at BestBuy.</td>
</tr>
<tr>
<td>Expired!</td>
</tr>
</tbody>
</table>

q

Deals

<table>
<thead>
<tr>
<th>product</th>
<th>seller</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>Amazon</td>
</tr>
<tr>
<td>iPhone</td>
<td>BestBuy</td>
</tr>
</tbody>
</table>

Deals

<table>
<thead>
<tr>
<th>product</th>
<th>seller</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCDs</td>
<td>Dell</td>
</tr>
<tr>
<td>TV</td>
<td>Amazon</td>
</tr>
</tbody>
</table>
Extractor Properties: Scope

- Mention attributes often appear in close proximity.
  - an extractor E has scope $\alpha$ iff any mention produced by E at most spans $\alpha$.

E extracts a deal if
(a) product & seller are within a window of at most 20 chars
(b) no “Expired!” is within 10 chars around the window.

**Hot Deals**
- Dec 11: All Dell LCDs $299.
- Dec 10: Sony TV $850 at Amazon.
- Dec 9: iPhone $199 at BestBuy. Expired!

E’s scope = 20 chars
Obtain Scope and Context

- Cyclex takes E, its scope & context

- Scope & context are provided by
  - who writes E
  - who knows how E works

- Even with loose scope and context
  - can still guarantee correctness
  - can still reduce runtime
  - tighter scope/context ➔ more recycling
Baseline Approach

Match

Overlapping Regions

Derive Regions

Copy Regions

Extraction Regions

Copy

Apply E
Matchers

- Consider 3 matchers (more can be added)
  - DN (Doing Nothing): 0 runtime, no overlapping regions
  - UD (Unix Diff): fast, find some overlapping regions
  - ST (Suffix Tree): relatively slow, all overlapping regions

- Trade off runtime vs. size of overlapping regions

- No matcher is always optimal
Select An Optimal Matcher: Adaptive and Cost-based Optimization

- Select Matcher
- Matchers: DN, UD, ST
- Cost Model

Look at past $k$ snapshots, $k$ is pre-defined.
Select An Optimal Matcher: Cost-based Selection

- Search for the optimal plan
  1. enumerate all possible plans
  2. use cost model to estimate cost of each plan
  3. select plan with minimal cost

Plan Space

Matchers
DN UD ST

DN
UD
ST

cost = 100
cost = 60
cost = 24
Cost Model

- Captures matching/extraction/copy times

\[
\text{matching} + \text{extraction} + \text{copy}
\]

- Captures text properties

\[
\hat{w}_{1,ex} \cdot m \cdot f \cdot l \cdot g
\]

- length (in byte)/page
- fraction of old pages
- fraction of new regions
- # of pages in \(S_4\)
- length of new regions on matched pages
Baseline Approach

Match

Overlapping Regions

Derive Regions

Copy Regions

Extraction Regions

Copy

Apply E
Interleave Matching, Extraction and Copy

Match

Overlapping Regions

Derive Regions

context scope

Hash Join

Copy

Copy Regions

Extraction Regions

Apply E
### Experimental Setup

#### Datasets

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>DBLife</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td># Data Sources</td>
<td>980</td>
<td>925</td>
</tr>
<tr>
<td># Snapshots</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Time between snapshots</td>
<td>1 day</td>
<td>21 days</td>
</tr>
<tr>
<td>Avg # Page per Snapshot</td>
<td>10155</td>
<td>3038</td>
</tr>
<tr>
<td>Avg Size per Snapshot</td>
<td>180M</td>
<td>35M</td>
</tr>
</tbody>
</table>

#### Extractors

<table>
<thead>
<tr>
<th></th>
<th>DBLife</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>researcher</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>affiliation</td>
<td>93</td>
<td>96</td>
</tr>
<tr>
<td>talk</td>
<td>400</td>
<td>250</td>
</tr>
<tr>
<td>actor</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>play</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>award</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scope α: 32 93 400 35 96 250

Context β: 3 7 10 3 4 10
Benefit of Recycling IE Results

- In 5 out of 6, outperformed extract-from-scratch by 50-90%
Importance of Optimization

- No matcher is uniformly optimal
Summary of Cyclex

- **Guarantee correctness**
  - model extractors using *scope and context*

- **Choose a good way to match pages**
  - adaptive optimization
  - cost-based decisions using *text specific* cost model

- **Efficiently interleave matching, copy and extraction**
  - a way to *scan data once*
My Dissertation Contributions

Developed efficient IE solutions over evolving text that

• recycle previous IE efforts
• guarantee correctness
• deal with large text corpora
• deal with large IE programs
• deal with complex learning-based IE programs

Results in [ICDE-08], [SIGMOD-09] and [TechReportA-10]
Real-world IE programs are complex

- Avatar@IBM: 25+ blackboxes
- DBLife@WISC: 45+ blackboxes
Delex [SIGMOD09]: Decompose and Recycle

How to coordinate recycling among "blackboxes"?

Assume a fast recycling plan?

Matcher: extractSentence
Matcher: extractProduct
Matcher: extractSeller
Matcher: extractDeal

product
seller

RDBMS query optimization
organize files for sequential access

exploit databases techniques!
Experiment Setup

- **Datasets**

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>DBLife</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td># Snapshots</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Time between snapshots</td>
<td>2 days</td>
<td>21 days</td>
</tr>
<tr>
<td>Avg # Page per Snapshot</td>
<td>10155</td>
<td>3038</td>
</tr>
<tr>
<td>Avg Size per Snapshot</td>
<td>180M</td>
<td>35M</td>
</tr>
</tbody>
</table>

- **IE Programs : Rule-based and Learning-based IE Programs**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>talk</td>
<td>chair</td>
<td>advise</td>
</tr>
<tr>
<td># of IE “Blackboxes”</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
Delex drastically cuts runtime of Cyclex by 45-71%
(See paper for more experiments)
Related Work

- **Early works on IE**
  - [Bikel UAI-97] [McCallum KDD-00]...
  - focus on improving accuracy

- **Recent works on IE**
  - [tutorial in KDD-06, SIGMOD-06] [Gravano et al, SIGMOD-06]...
  - consider developing efficient IE
  - do not consider evolving text corpora

- **Evolving text data**
  - [McCann VLDB05] [Lim WWW03]
  - consider problems other than IE

- **Incremental view maintenance**
  - [Gupta&Mumick][Widom et al, SIGMOD95]...
  - only consider relational operators
  - assume changes to the inputs are available
My Other Contributions

- **Web community systems**
  - DBLife [DE Bulletin-06, VLDB-07a, CIDR-09]

- **Text mining in cloud**
  - scalable mining algorithms on MapReduce clusters [TechReportB-10]

- **Biology sequences mining**
  - incorporate domain knowledge into mining algorithms [KDD-03, Bioinformatics-04]
  - interpret complex mining models with human-understandable rules [SFUThesis]
Future Work

● Maintain IE and information integration (II)

● Develop/Maintain IE/II over large clusters
  – build efficient, scalable and robust IE/II
  – declaratively develop and automatically optimize IE/II

● Develop user-friendly and efficient data analysis tools
  – analyze non-structured data, e.g., text and biology sequences
  – let non-database users such as biologists ask query easily and obtain answers efficiently
Conclusions

- IE over evolving text is increasingly important!

- Developed solutions that
  - recycle previous IE to reduce runtime
  - guarantee correctness
  - deal with large text corpora / large IE programs
  - deal with complex learning-based IE programs

- Database techniques are increasingly critical for developing efficient IE solutions
  - in collaboration with other communities: NLP, AI, Web, IR, KDD, HCI