Optimizing Complex Extraction Programs over Evolving Text Data

Fei Chen¹, Byron Gao², An Hai Doan¹, Jun Yang³, Raghu Ramakrishnan⁴

¹University of Wisconsin-Madison
²Texas State University-San Marcos
³Duke University
⁴Yahoo! Research
Many solutions in database/Web/AI communities with significant progress

But most solutions have considered only *static text corpora*
Evolving Text Corpora Are Pervasive

- **Impliance @ IBM**
  - find the latest information from enterprise intranets

- **IWP@Univ. of Washington and YAGO@MPI**
  - keep extracted knowledge consistent with the Wikipedia pages

- **DBLife@Univ. of Wisconsin**
  - monitor community information
Group Meeting Schedule
Jun 14: Meet in CS 105 at 2pm.

Group Meeting Schedule
Jun 21: We’ll discuss CIM and IR in room CS 310 at 4pm.
Jun 14: Meet in CS 105 at 2pm.

Group Meeting Schedule
Jun 28: Seminar in CS 354 at 4pm.
Jun 21: We’ll discuss CIM and IR in room CS 310 at 4pm.
Jun 14: Meet in CS 105 at 2pm.
Group Meeting Schedule

Jun 14: Meet in CS 105 at 2pm.

Jun 21: We’ll discuss CIM and IR in room CS 310 at 4pm.

**Reuser**

**Matcher**

**Group Meeting Schedule**

Jun 14: Meet in CS 105 at 2pm.

Jun 21: We’ll discuss CIM and IR in room CS 310 at 4pm.

**room** | **time**
--- | ---
CS 105 | 2pm
CS 310 | 4pm
Cyclex[ICDE08]: Properties of Extractors for Correct Reuse

- **Scope**: max length of any mention extracted by E

- **Context**: length of “text windows” surrounding a mention
  - E only exams the text windows to extract the mention

Example: E extracts telephone numbers using regular expression

```
“be reached at \d{7}”
```

context = 14 chars  scope = 7 chars
Limitations of Cyclex [ICDE08]

- Cyclex treats an IE program as a blackbox
- Real-world IE programs are complex
  - Avatar: 25+ blackboxes
  - DBLife: 45+ blackboxes
Delex: Decompose and Recycle

- Exploit the composition nature of IE programs
- Delex cuts the runtime of Cyclex by 50-71%
Delex on Consecutive Snapshots

Snapshot n

Matchers
Cost Model
Select a good reuse plan
Execute the plan
Capture IE results

Snapshot n+1

Matchers
Cost Model
Select a good reuse plan
Execute the plan
Capture IE results

Captured results of snapshot n
Captured results of snapshot n+1
A Baseline Solution of Capturing Results for Multi-Blackbox IE Programs

- Capturing results in Cyclex

- A baseline solution: applying the solution of Cyclex to each blackbox
The Baseline Solution May Miss Mentions

**Midwest DB Courses:**
- CS101 (Wisc)
- CS301 (Illinois)

**Midwest DB Courses**
- CS101 (Wisc)
- CS301 (Illinois)

Year 2009

Locations(location)
- exLocation
  - rmTitle

Midwest

Locations(location)
- exLocation
  - rmTitle

no mention is copied

no mention is copied

missed!
Capture and Store IE Results in Delex

Midwest DB Courses:
CS101 (Wisc)
CS301 (Illinois)

Locations(location)

I_{exLocation} (q)  I_{exLocation} (r)  ...
O_{exLocation} (q)  O_{exLocation} (r)  ...

Wisc
Illinoins

"CS101 … Illinois)"
Delex on Consecutive Snapshots

Snapshot n
- Matchers
- Cost Model
  - Select a good reuse plan
  - Execute the plan
  - Capture IE results

Snapshot n+1
- Matchers
- Cost Model
  - Select a good reuse plan
  - Execute the plan
  - Capture IE results

Captured Results
Challenge in Efficiently Reuse Captured Results

\[ S_n q_2 \]
\[ \cdots I_U (q_1) \cdots I_U (q_2) \cdots \]
\[ O_U \]
\[ \cdots O_U (q_1) \cdots O_U (q_2) \cdots \]
\[ I_V \]
\[ \cdots I_V (q_1) \cdots I_V (q_2) \cdots \]
\[ O_V \]
\[ \cdots O_V (q_1) \cdots O_V (q_2) \cdots \]
Overall Processing Algorithm

\[
\begin{align*}
&I_U(q_1) \quad I_U(q_2) \quad \cdots \\
&O_U(q_1) \quad O_U(q_2) \quad \cdots \\
&I_V(q_1) \quad I_V(q_2) \quad \cdots \\
&O_V(q_1) \quad O_V(q_2) \quad \cdots
\end{align*}
\]
Delex on Consecutive Snapshots

Snapshot n

Matchers

Cost Model

Select a good reuse plan

Execute the plan

Capture IE results

Captured Results

Snapshot n+1

Matchers

Cost Model

Select a good reuse plan

Execute the plan

Capture IE results

Captured Results
Find a Good Reuse Plan

- **Plan space**
  - assign a matcher to each IE blackbox
  - # of plans is **exponential** in # of IE blackboxes

- **Use a text-specific cost model to estimate the completion time of each plan**

- **Searching for good plans**
  - optimization is **not “decomposable”**
  - a greedy solution that efficiently finds a good plan
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>
Runtime Comparison

- Delex drastically cuts runtime of cyclex by 50-71%
  (See paper for more experiments)
Related Work

• **IE over evolving text data**
  - [Doan et al, ICDE-08]
  - only considers a single IE blackbox

• **Optimizing IE programs**
  - [Gravano et al, SIGMOD-06] [Gravano et al, ICDE-07] [Doan et al, VLDB-07] [Reiss et al, ICDE-08]
  - only consider static text corpora

• **Incremental View Maintenance**
  - [Gupta&Mumick][&Widom et al, SIGMOD-95][Garcia-Molina&Widom et al, VLDB-91]…
  - only consider relational operators with well defined semantics
  - assume that changes to the inputs are readily available
Conclusion and Future Work

• First in-depth solution to optimizing complex IE over evolving text

• Defined challenges and provided initial solutions
  – capture intermediate IE results for correct reuse
  – efficiently coordinate matching, extraction, and copying for multiple IE blackboxes
  – cost-based decisions in choosing a good reuse plan

• Future work
  – reuse across URLs
  – handle extractors that extract mentions across multiple pages