Web Information Extraction

Mining Web Pages for Data Records
by Liu, Grossman, Zhai

(Paper #8)
Presentation by Paul Bohunsky (0025058)
Main goal

- Exploits Web page structure to mine contiguous and non-contiguous data records.
Main goal

- Exploits Web page structure to mine contiguous and non-contiguous data records.
- uses a string-matching algorithm.
Main goal

- Exploits Web page structure to mine contiguous and non-contiguous data records.
- Uses a string-matching algorithm.
- Stand-alone system.
1. Build an HTML tag tree.
1. Build an HTML tag tree.

2. Mine all data regions using *edit string-matching algorithm*.
Technique Overview (3)

1. Build an HTML tag tree.
3. Identify data records from data regions.
A group of similar Data Records (containing similar objects) appearing in a contiguous region of a page is forming a Data Region.
Data Region

- A group of similar Data Records (containing similar objects) appearing in a contiguous region of a page is forming a Data Region
- Collection of two or more Generalized Nodes.
HTML Tag Tree

(a)

<table>
<thead>
<tr>
<th>Brand X Notebook RX300</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Buy new:</strong> $1,299</td>
<td></td>
</tr>
<tr>
<td><strong>Customer Rating:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>(what’s this?)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>600 MHz, G3, 256 MB SRAM, 40 GB Hard Disk, 24x CD-ROM, Microsoft Internet Explorer.</strong></td>
<td></td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Brand X Notebook P6000</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Buy new:</strong> $1,699</td>
<td></td>
</tr>
<tr>
<td><strong>Customer Rating:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>(what’s this?)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>667 MHz, G4, 528 MB SDRAM, 80 GB Hard Disk, 24x CD-ROM, Microsoft Word.</strong></td>
<td></td>
</tr>
</tbody>
</table>
Generalized Nodes =
Node Combination

• They have the same parent.
Generalized Nodes = Node Combination

- They have the same parent.
- They have the same length.
Generalized Nodes = Node Combination

- They have the same parent.
- They have the same length.
- They are adjacent.
Generalized Nodes = Node Combination

- They have the same parent.
- They have the same length.
- They are adjacent.
- The normalized edit distance is less than a fixed threshold.
Generalized Nodes = Node Combination

- They have the same parent.
- They have the same length.
- They are adjacent.
- The normalized edit distance is less than a fixed threshold.
- Tag string == String of subtree.
Determine Data Regions

- Where does the first generalized node start?
Determine Data Regions

- Where does the first generalized node start?
- How many tag nodes (components) are in the generalized node?
Example

tag string for second <tr> node under <table> would be:

<tr td td ... td td>
Finding Data Regions (1)

- Depth-first search (recursively)
Finding Data Regions (1)

- Depth-first search (recursively).
- String comparison with adjacent nodes.
  (1-2), (2-3), ..., (1-2,3-4),...
  (1-2-3, 4-5-6), ...
Finding Data Regions (1)

- Depth-first search (recursively).
- String comparison with adjacent nodes.
  (1-2), (2-3), ..., (1-2,3-4),..., 
  (1-2-3, 4-5-6), ...
- --> Generalized Nodes.
Finding Data Regions (2)

- If a higher data region covers a lower one, only the higher (and its Generalized Nodes) is reported.
Finding Data Regions (2)

- If a higher data region covers a lower one, only the higher (and its Generalized Nodes) is reported.
- If s1, s2, ..., sn strings are similar, report only the smallest Generalized Node.
Identifying Data Records (1)

- If child nodes of GN are similar and GN is not a <tr> then each child node of GN is a Data Record.
If all child nodes of GN are similar and each node has same number of children, child nodes of every node in GN form a non-contiguous object description.

<table>
<thead>
<tr>
<th>Row 1</th>
<th>Name 1</th>
<th>Name 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 2</td>
<td>Description 1</td>
<td>Description 2</td>
</tr>
<tr>
<td>Row 3</td>
<td>Name 3</td>
<td>Name 4</td>
</tr>
<tr>
<td>Row 4</td>
<td>Description 3</td>
<td>Description 4</td>
</tr>
</tbody>
</table>
Summary

- Extracts Data Records if they are represented with table- and/or form-related tags only
Summary

- Extracts Data Records if they are represented with table- and/or form-related tags only.
- Can handle non-contiguous data descriptions.
Summary

- Extracts Data Records if they are represented with table- and/or form-related tags only.
- Can handle non-contiguous data descriptions.