Automating the Extraction of Genealogical Information from the Web

GeneTIQS

Troy Walker & David W. Embley
Family History Technology Conference
March 25, 2004

Research funded by NSF grant #IIS-0083127
Genealogical Information on the Web

- **Hundreds of thousands of sites**
  - Some professional (Ancestry.com, Familysearch.org)
  - Mostly hobbyist (203,200 indexed by Cyndislist.com)

- **Search engines**
  - “Walker genealogy” on Google: 199,000 results
  - 1 page/minute = 5 months to go through

- **Why not enlist the help of a computer?**
Problems

- No standard way of presenting data
- Sites have differing schemas
- Web pages change
- New pages continuously come on line
GeneTIQS

- Based on work done by BYU DEG
- Able to extract from:
  - Single-record documents
  - Simple multiple-record documents
  - Complex multiple-record documents
- Robust to changes in pages
- Immediately works for new pages
Person Ontology
Value Matchers
Record Separation

- Separating data related to each person
- Previous technique
  - Combines many heuristics
  - Has problems
    - Assumes multiple records
    - Must be simple separation
Adeline Theresa Marie Adams

Date of Birth: February 9, 1875
Place of Birth: Plymouth, Devon, England
Father: George Absalom Adams
Mother: Margaret Mary Ann Hooper
Date of Marriage: September 9, 1897
Place of Marriage: Unknown
Spouse: George Alfred Bulmer-Bird
Date of Death: July 13, 1953
Place of Death: Manningtree, Essex, England
Children:
- Adeline Dorothy
- Cecil George
- Constance Margaret

Generation Number: 4
<table>
<thead>
<tr>
<th>Name: Blaine Wright ANDERSON / Carolyn TENSMEYER</th>
<th>Microsoft Internet Explorer</th>
<th>Address: <a href="http://www.sanbouveau.com/boice/Kempf0000425.html#456">http://www.sanbouveau.com/boice/Kempf0000425.html#456</a></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Born:</strong> 24 Oct 1880</td>
<td><strong>Died:</strong> 25 Apr 1896</td>
<td>Spouses:</td>
</tr>
<tr>
<td><strong>Name:</strong> Don Carlos WALKER</td>
<td></td>
<td><strong>at:</strong> Taylor, Apache, AZ</td>
</tr>
<tr>
<td><strong>Born:</strong> 21 May 1883</td>
<td><strong>Died:</strong> 27 Jan 1972</td>
<td><strong>at:</strong> Darrango, La Plata, CO</td>
</tr>
<tr>
<td><strong>Name:</strong> Ezra Francis WALKER</td>
<td></td>
<td><strong>at:</strong> Taylor, Apache, AZ</td>
</tr>
<tr>
<td><strong>Born:</strong> 27 Sep 1885</td>
<td><strong>Died:</strong> 19 Sep 1952</td>
<td><strong>at:</strong></td>
</tr>
<tr>
<td><strong>Name:</strong> Estelle WALKER</td>
<td></td>
<td><strong>at:</strong> Taylor, Apache, AZ</td>
</tr>
<tr>
<td><strong>Born:</strong> 5 Aug 1887</td>
<td><strong>Died:</strong> 23 Dec 1980</td>
<td><strong>at:</strong> Taylor, Apache, AZ</td>
</tr>
<tr>
<td><strong>Name:</strong> Josephine WALKER</td>
<td></td>
<td><strong>at:</strong> Taylor, Apache, AZ</td>
</tr>
<tr>
<td><strong>Born:</strong> 15 Apr 1890</td>
<td><strong>Died:</strong> 15 Apr 1953</td>
<td><strong>at:</strong></td>
</tr>
<tr>
<td><strong>Name:</strong> Walter Noah WALKER</td>
<td></td>
<td><strong>at:</strong> Taylor, Apache, AZ</td>
</tr>
<tr>
<td><strong>Born:</strong> 18 Aug 1892</td>
<td><strong>Died:</strong> 13 Feb 1940</td>
<td><strong>at:</strong> Taylor, Apache, AZ</td>
</tr>
<tr>
<td><strong>Name:</strong> Lavena WALKER</td>
<td></td>
<td><strong>at:</strong> Taylor, Navajo, AZ</td>
</tr>
<tr>
<td><strong>Born:</strong> 25 Oct 1894</td>
<td><strong>Died:</strong> 12 Nov 1965</td>
<td><strong>at:</strong></td>
</tr>
</tbody>
</table>
Vector Space Modeling

- Ontology Vector
- Compare to candidate records
  - Cosine measure
    \[ v_1 \cdot v_2 \]
  - Magnitude measure
    \[ |v_1| \]
Ontology Vector

{ 0.8, 0.99, 0.95, 0.9, 0.6, 0.5, 0.6, 3.0}
Vector Space Modeling

```html
<!DOCTYPE...>
<html>
  <head>
    ...
  </head>
  <body>
    <div>
      ...header...
    </div>
    <div>
      ...
    </div>
    <div>
      ...
    </div>
    <div>
      ...
    </div>
  </body>
</html>
```
Improvements

- **Differing schemas**
  - Low cosine measures
  - Discarded data
  - Prune dimensions
    - \{0.8, 0.99, 0.95, 0.9, 0.6, 0.5, 0.6, 3.0\}
    - \{0.0, 141.0, 89.0, 76.0, 0.0, 0.0, 48.0, 23.0\}

- **Richness of data in single-record documents**
  - High magnitude measure
  - Higher magnitude to split documents
Demonstration
Presenting Results

Person Query Results

<table>
<thead>
<tr>
<th>Gender</th>
<th>Name</th>
<th>Birth Location</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
<th>Death Location</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
<th>Christening Location</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
<th>Burial Location</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
<th>Marriage</th>
<th>Relationship</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adeline Theresa Marie Adams</td>
<td>Plymouth, Devon, England</td>
<td>8</td>
<td>February</td>
<td>1875</td>
<td>Manningtree, Essex, England</td>
<td>13</td>
<td>July</td>
<td>1953</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lillian Holmes</td>
<td>Manningtree, Essex, England</td>
<td>14</td>
<td>April</td>
<td>1859</td>
<td>Manningtree, Essex, England</td>
<td>30</td>
<td>October</td>
<td>1913</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>William Smallwood</td>
<td>Stafford Co, VA</td>
<td>23</td>
<td>JUL</td>
<td>1725</td>
<td>Edgerfield Co, SC</td>
<td>3</td>
<td></td>
<td>1782</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>John Barrett</td>
<td>Ohio</td>
<td>1841</td>
<td></td>
<td>1841</td>
<td></td>
<td>1862</td>
<td></td>
<td>1862</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thomas Walker</td>
<td>22 Feb</td>
<td>1765</td>
<td>22 Feb</td>
<td>1799</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thomas H. Sheldon</td>
<td></td>
<td>1818</td>
<td></td>
<td>1854</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Martha Unics</td>
<td>1824</td>
<td></td>
<td></td>
<td>1812</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Courtland Philip Livingston Butler</td>
<td></td>
<td>1813</td>
<td></td>
<td>1891</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elizabeth Slade Pierce</td>
<td></td>
<td>1822</td>
<td></td>
<td>1901</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Robert Emmet Sheldon</td>
<td></td>
<td>1845</td>
<td></td>
<td>1917</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Preliminary Results

- Semi-structured Text
  - 10 single-record documents
  - 3 simple documents containing 268 records
  - 3 complex documents containing 266 records
- Precision and recall for record separation
# Record Separation

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>100%</td>
<td>94.1%</td>
</tr>
<tr>
<td>Simple</td>
<td>94.7%</td>
<td>97.3%</td>
</tr>
<tr>
<td>Complex</td>
<td>88.3%</td>
<td>93.6%</td>
</tr>
</tbody>
</table>
Conclusion

- Integrate, build on previous DEG work
- Accurate record separation
  - Average recall: 94.3%
  - Average precision: 95.0%
- Ontology based
  - Robust to changes in pages
  - Immediately works with new pages