Retrieving a Sorted List of Hundreds of Closest Relatives from FamilySearch Family Tree in Seconds

Family History Technology Workshop
Brigham Young University
March 20, 2014

Ben Baker
bakerb@familysearch.org
Weighted Relationship Distance

\[ \text{WRD}_{(g, c, m)} = \alpha(|g| + 1) e^{\beta c} e^{\gamma m} \]

- \( g \) – Generational or “vertical” distance
  Number of generations from base person
- \( c \) – Collateral or “horizontal” distance
  Minimum generations to a closest common ancestor
- \( m \) – Marriage distance
  Number of marriages between base person
- \( \alpha, \beta, \gamma \) – Weighting factors to control growth rates

2013 Family History Technology Workshop – Beyond the Relationship Calculator
Using a Weighted Relationship Distance Metric to Prioritize, Categorize and Visualize Relatives
Sample Relatives and WRD values

- Base Person: WRD (0, 0, 0) = 1.0
- Father: WRD (1, 0, 0) = 2.0
- Grandfather: WRD (2, 0, 0) = 3.0
- Grandmother: WRD (2, 0, 0) = 3.0
- Son: WRD (-1, 0, 0) = 3.52
- Niece: WRD (-1, 1, 0) = 9.57
- Sister-in-law: WRD (0, 1, 1) = 11.25
- Brother-in-law: WRD (0, 1, 1) = 11.25
- Sister-in-law: WRD (0, 1, 2) = 46.53
- Wife: WRD (0, 0, 1) = 4.14
- Niece: WRD (-1, 1, 1) = 39.59
- Aunt: WRD (1, 1, 0) = 5.44
- Uncle: WRD (1, 1, 1) = 22.49
- 1st Cousin: WRD (0, 2, 0) = 7.39
- Cousin: WRD (0, 2, 0) = 7.39
- Mother: WRD (1, 0, 0) = 2.0
- Mother-in-law: WRD (1, 0, 1) = 8.27
- Sister: WRD (1, 1, 0) = 5.44
- Aunt: WRD (1, 1, 1) = 22.49
- Brother: WRD (0, 1, 0) = 2.72
- Niece: WRD (-1, 1, 1) = 39.59
- Husband of Grandmother: WRD (2, 0, 0.5) = 6.10
- Father-in-law: WRD (1, 0, 1) = 8.27
- Sister: WRD (1, 1, 0) = 5.44
- Aunt: WRD (1, 1, 1) = 22.49
- Uncle: WRD (1, 1, 1) = 22.49
- Aunt: WRD (1, 1, 0) = 5.44
- 1st Cousin: WRD (0, 2, 0) = 7.39
- Cousin: WRD (0, 2, 0) = 7.39
- Mother: WRD (1, 0, 0) = 2.0
- Mother-in-law: WRD (1, 0, 1) = 8.27
- Sister: WRD (1, 1, 0) = 5.44
- Aunt: WRD (1, 1, 1) = 22.49
- Uncle: WRD (1, 1, 1) = 22.49
- Aunt: WRD (1, 1, 0) = 5.44
- 1st Cousin: WRD (0, 2, 0) = 7.39
- Cousin: WRD (0, 2, 0) = 7.39
- Mother: WRD (1, 0, 0) = 2.0
- Mother-in-law: WRD (1, 0, 1) = 8.27
- Sister: WRD (1, 1, 0) = 5.44
- Aunt: WRD (1, 1, 1) = 22.49
<table>
<thead>
<tr>
<th>Relative Description</th>
<th>g</th>
<th>c</th>
<th>m</th>
<th>WRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Person</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>Father/Mother</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2.0</td>
</tr>
<tr>
<td>Brother</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2.72</td>
</tr>
<tr>
<td>Grandfather/Grandmother</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3.0</td>
</tr>
<tr>
<td>Son</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>3.52</td>
</tr>
<tr>
<td>Wife</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4.14</td>
</tr>
<tr>
<td>Aunt/Uncle (sibling of parent)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5.44</td>
</tr>
<tr>
<td>Husband of Grandmother</td>
<td>2</td>
<td>0</td>
<td>0.5</td>
<td>6.10</td>
</tr>
<tr>
<td>1st cousin</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>7.39</td>
</tr>
<tr>
<td>Father/mother-in-law</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>8.27</td>
</tr>
<tr>
<td>Niece (daughter of brother)</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>9.57</td>
</tr>
<tr>
<td>Brother/sister-in-law (spouse of sibling or sibling of spouse)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>11.25</td>
</tr>
<tr>
<td>Daughter-in-law</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>14.56</td>
</tr>
<tr>
<td>Aunt/Uncle (spouse of sibling of parent)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>22.49</td>
</tr>
<tr>
<td>Niece (daughter of spouse’s brother)</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>39.59</td>
</tr>
<tr>
<td>Sister-in-law (wife of spouse’s brother)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>46.53</td>
</tr>
</tbody>
</table>
Extreme Example
Relationships Between Benjamin Baker and Laura Welch (Bush)
Relationship Calculator Deficiencies

1. A relationship calculation must be initiated between two persons in an ad hoc manner and repeated for a different set of two persons.
2. It is not possible to sort a list of arbitrarily related persons by closeness.
3. Relationship calculations through marriages are not possible.
4. Data to perform the relationship calculations must be pre-calculated and stored to be performant enough for on-demand calculation in an enterprise system.
Apache Cassandra-Based FamilySearch Family Tree

Data is striped across n number of machines

VS.
Proposed System

RESTful web service endpoint added to 5-node Cassandra based test system

Method / Description

GET relatives/{id}

Retrieve the closest relatives to a person up to a specified maximum weighted relationship distance.

Parameters

double maxDistance – Optional query parameter (default 10.0) to specify the maximum distance to return relatives of the person up to.

Returns

A list of RelativePerson objects, sorted by increasing distance

Future methods planned to retrieve closest common ancestor and closest relation between two people.
RelativePerson Object

{
    "name": "George Dean Cockle",
    "id": "KWJX-VMC",
    "lifespan": "1901 - 1970",
    "fatherIds": [
        "K2V7-PV5"
    ],
    "motherIds": [
        "K2V7-P92"
    ],
    "spouseIds": [
        "MM8P-MGS",
        "KWBB-7G9"
    ],
    "familyName": "Cockle",
    "givenName": "George Dean",
    "gender": "MALE",
    "childIds": [
    ],
    "relative": {
        "id": "K2V7-PV5",
        "name": "John Cockle",
        "lifespan": "1852 - 1921",
        "relativeRole": "CHILD"
    },
    "weightedRelationshipDistance": {
        "starRanking": 8,
        "simpleRelationshipDistance": "3",
        "weightedRelationshipDistance": "8.155",
        "generationDistance": 2,
        "collateralDistance": 1,
        "marriageDistance": "0"
    },
    "relativeDescription": "Great Uncle"
}


Performance Results

<table>
<thead>
<tr>
<th>Maximum WRD</th>
<th>Num Persons</th>
<th>Mean Execution Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>43</td>
<td>1.95s</td>
</tr>
<tr>
<td>10.0</td>
<td>167</td>
<td>22.1s</td>
</tr>
<tr>
<td>15.0</td>
<td>554</td>
<td>89.8s</td>
</tr>
</tbody>
</table>

Performance tuning is still necessary to make this service more useful for production use.

For comparison purposes, the average time to retrieve data for a pedigree containing 10-15 persons in FamilySearch Family Tree is typically about 3-5 seconds. The Eureka team has shown sub-second pedigree load time involving the same number of persons and loading up to 12 generations in several seconds.
Live Demo
Future Applications Enabled

- Identifying the closest relatives where historical record hints have been identified but not attached yet as sources.
- Promoting e-mail campaigns to point out what others have added to your relatives such as new photos, sources, stories, etc. to draw users back to the site.
- Easily identifying end of line relatives and likely places for successful descendancy research.
- Facilitate LDS temple work for closest relatives first and sharing more distantly related people with others.
- Producing to-do lists sorted by closeness of relation on any task a user may want to undertake (Ex. fixing data anomalies, merging possible duplicates, providing missing data, etc.)
- Automatic watching of relatives via e-mail alerts based on closeness.
- Applications across a set of users (Ex. closeness of relation to the user for all LDS temple submissions or photo uploads on FamilySearch)
- Sorting items such as memories, watched persons, LDS temple reservation lists, etc. in order of those closest to the user.
- Pointing out relatives who have been identified as prominent or have participated in significant events.
- More . . .
Next Steps

• Full transition from Oracle to Cassandra expected to take a while, but expect one-way synchronization sooner
• Enables read-only operations on Family Tree data, including this work
• Intend to petition to utilize this work in production before transition is complete
• May implement a smaller-scale solution in the current system to prove value of WRD metric over current scope of interest service
Additional Q & A