MEDD: An Approximate Matching Technology for Database Searching, Linking, and De-Duplicating

Arthur Goldberg and Andrew Borthwick

Practice-Oriented Paper

Executive Summary

When you need to combine multiple, error-filled data feeds into a single, highly accurate database, the hardest problem is matching corresponding records. How do you match, for instance, "Thomas J. Hanks" with "Tom Hank" or “International Business Machines” with “Intl. Bus. Mach.”? We present an innovative, accurate system that employs a powerful, patent-pending, machine learning technique to determine the probability that two database records correspond to the same person or company.

We start by showing why record matching is such a difficult problem and describe the basics of the record matching process. As an example, we discuss the New York City Department of Health, where we removed 300,000 duplicate records from a 2.1 million record children’s health database.

MEDD is built around “comparison functions”. Comparison functions check whether a pair of records has a certain matching or non-matching characteristic. Examples include “First names match”, “First names match using the ‘Soundex’ phoneticization technique”, or “Birthday does not match”.

MEDD uses a training process called “maximum entropy modeling” to infer the relative importance of the different comparison functions from a small set of record-pairs which have been hand-marked as “same” or “different”. Out of this process comes a “weight” which is assigned to each feature.

At runtime, MEDD operates as a function which takes a set of fields (a “search record”) as an input and returns a list of database ID’s which might match the search record. The ID’s are ranked by a probability of match which is computed by MEDD’s weighted comparison functions.
Approximate Record Matching
- Record matching tasks
  - Remove duplicates from a database
  - Link multiple databases
  - Search a database for a record
- Matching difficulties
  - No unique IDs
    - Some databases prohibit SSNs
  - Incorrectly entered data
    - Borthwick vs. Borthwich
  - Time-varying data
    - Address changes
  - Inconsistently used identifying data
    - Andrew vs. Andy

Matching Catastrophes
- NYC Department of Health Child DB
  - 1.4M children duplicated into 2.1M records
- Removing felons from Florida's voter roles
  - Some counties purged non-felons.
  - Some counties did no purge because of list's inaccuracies
- Wall street business data
  - Two clerks work full time matching by hand

MEDD Matches Healthcare Data
- Client: NYC Department of Health
- Projects
  1. Remove immunization database duplicates
     - Prevent over and under immunization
  2. Link immunization and lead-exposure test databases
     - Enable caseworkers to address both under-immunization and lead exposure when visiting clients

NYC Immunization Database
- Parameters
  - NYC birth cohort 122,000
  - Over 2M records
  - Monthly updates from 1,100+ institutions and providers
    - Up to 100,000 patients
    - Up to 200,000 immunization events
- Before MEDD: 3 records for every 2 kids
- Strict criteria for automatic merging
- In 1996 clerks manually de-duplicated
  - 260,000 record pairs
  - 1,700 person-hours

MEDD De-Duplicates NYC Immunization Database
- Work in 1999-2000

<table>
<thead>
<tr>
<th>Birth year</th>
<th>Records</th>
<th>Dupes removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>203,389</td>
<td>25,553</td>
</tr>
<tr>
<td>1997</td>
<td>216,336</td>
<td>34,773</td>
</tr>
<tr>
<td>1998</td>
<td>208,315</td>
<td>47,830</td>
</tr>
<tr>
<td>1999</td>
<td>157,946</td>
<td>42,228</td>
</tr>
<tr>
<td>TOTAL</td>
<td>785,986</td>
<td>150,384</td>
</tr>
</tbody>
</table>
**MEDD Links Two Databases**

- Databases
  - Immunization
  - Lead exposure

- Synergy between the two programs
  - The same kids can be under-immunized and missing a lead screening test
  - Both databases cover all NYC children

- Finish in early 2001

---

**NYC MEDD/MCI System**

- Information about every child in either database is stored in a MEDD-based Master Child Index (MEDD/MCI)
- Each system can retrieve data from the other by finding corresponding IDs in the MEDD/MCI

---

**MEDD/MCI Record Matching**

- Remove duplicates
- Connect immunization and lead exposure children
- Determine whether incoming records are already in MCI
- Periodically scan MCI for residual duplicates

---

**NYC DOH’s Benefits from MEDD**

**Savings**
- Automatically removed 200,000 records in ’99-’00
  - Original process would have required hand-examining at least 600,000 record-pairs
  - Cost of 2 person-years
- To summer ’01, almost 600,000 records removed

**Improvements**
- Matching incoming records prevents creation of duplicates
- Enabled linkage of immunization and lead databases
- Old process was much less accurate
  - Error rate of a typical clerk is over 1%
  - Clerks only reviewed very similar records. Many “tricky” matches were never reviewed
- DOH accepting “noisy” data feeds (billing feeds from HMO’s, forms filled out in doctor offices)

---

**Production Matching Basics**

**Input** Search record

**Blocking**
- Find thousands of possible matches

**Match decision making**
- For each possible match
  - Evaluate many comparison functions against search record
  - Combine comparison functions by weight to produce match probability

**Output** IDs and probabilities of likely matches

---

**Production Matching**

- Search Record
- Blocking
- Many Possible Matches
- Maximum Entropy Matching
- Match Probabilities of Likely Matches
- Non-Match
- Intermediate
- Match
- Human Review
### Database of Children

- Do first names match?
- Do first names match approximately using "phonetic matches" such as Soundex, edit-distance, NYSIIS, or Jaro-Winkler?
- Do uncommon first names match?
- Do we have an indicator that the child is part of a multiple birth?
- Do Medicaid numbers match or mismatch?
- Do birthdays match?

### Database of Businesses

- How many words in the name match?
- Can the names be converted to the same abbreviation?
- Are the names the same after translating foreign words to English?
- Do country, phone number, or street address match?

### Complex Comparison Functions

#### Adapt to database quirks

**Child medical database example**

- **HMO XYZ sends Day of Birth = “1”**
  - Birthday = July 1, 1998 not July 15, 1998

**A special comparison function**

- If Provider = “HMO XYZ”
  - AND Day of Birth = 1
  - AND dates differs only on day of birth
  - THEN Match

### Customized with Java

**Java-based Comparison Functions**

- Simple first-name Soundex comparison function:

  ```java
  feature firstNameSoundexMatch {
      match equals(soundex(FIRST_NAME));
  }
  ``

- Comparison function for the HMO example on the previous slide:

  ```java
  feature HMOXYZandFirstOfMonth {
      match ((q.FACILITY_ID == "XYZ" && q.DOB.getDay() == 1) ||
             (m.FACILITY_ID == "XYZ" && m.DOB.getDay() == 1) &&
             q.DOB.getMonth() == m.DOB.getMonth() &&
             q.DOB.getYear() == m.DOB.getYear());
  }
  ```

### Maximum Entropy Matching Math

- The probability a pair of records match

  $$
  \text{MatchProduct} = \text{product of weights of all comparison functions predicting Match for the pair}
  $$

  $$
  \text{No-MatchProduct} = \text{product of weights of all comparison functions predicting No-Match for the pair}
  $$

  $$
  \text{MatchProduct} + \text{No-MatchProduct}
  $$

### MEDD Decides Match

**99.5% Confidence**

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Record</th>
<th>Match?</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surname</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOB</td>
<td>4/28/97</td>
<td></td>
<td>1.138</td>
</tr>
<tr>
<td>Street</td>
<td>4528 3rd Ave</td>
<td></td>
<td>4.342</td>
</tr>
<tr>
<td>City</td>
<td>Bronx</td>
<td></td>
<td>1.103</td>
</tr>
<tr>
<td>State</td>
<td>NY</td>
<td></td>
<td>3.013</td>
</tr>
<tr>
<td>Zip</td>
<td>10462</td>
<td></td>
<td>2.130</td>
</tr>
<tr>
<td>Phone</td>
<td>718-123-4567</td>
<td>No-match</td>
<td>6.087</td>
</tr>
<tr>
<td>Med Rec Number</td>
<td>11856437503</td>
<td>Match</td>
<td></td>
</tr>
</tbody>
</table>

Match product = 587.2
No-Match product = 2.9

$$
587.2 \times 2.9 = 0.995
$$
**MEDD Decides No-match**

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Record 1</th>
<th>Record 2</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last name</td>
<td>Lopez</td>
<td>Lopez</td>
<td>1.153</td>
</tr>
<tr>
<td>First name</td>
<td>Lopez</td>
<td>Lopez</td>
<td></td>
</tr>
<tr>
<td>Soundex First name</td>
<td>Girl</td>
<td>Susan</td>
<td></td>
</tr>
<tr>
<td>DOB</td>
<td>1/11/97</td>
<td>1/12/97</td>
<td>No-match</td>
</tr>
<tr>
<td>Street</td>
<td>987 Cornell</td>
<td>456 Park</td>
<td>No-match</td>
</tr>
<tr>
<td>City</td>
<td>Brooklyn</td>
<td>Brooklyn</td>
<td>1.103</td>
</tr>
<tr>
<td>State</td>
<td>NY</td>
<td>NY</td>
<td></td>
</tr>
<tr>
<td>Zip</td>
<td>11211</td>
<td>11211</td>
<td>Match</td>
</tr>
<tr>
<td>Phone</td>
<td>718-123-4567</td>
<td>718-234-5678</td>
<td>No-match</td>
</tr>
<tr>
<td>Med Rec Number</td>
<td>1001002</td>
<td>567435</td>
<td></td>
</tr>
</tbody>
</table>

MatchProduct = 3.8
No-MatchProduct = 181.1
\[ \frac{3.8}{181.1} = 0.021 \]

---

**MEDD Deployment**

**Principles of Maximum Entropy**

**How are weights determined?**
- Input record pairs marked Match or No-match.
- Weights selected so model predicts average probability of match for each comparison function equal to average probability for that comparison function in training data.

- **Name**
  - Probability records match given that name matches = 2/3
- **Phone**
  - Probability records match given that Phone matches = 7/9

---

**Matching with MEDD**

**Threshold Probabilities**

**Tradeoff Human Review Against Accuracy**

---

**Technical Information**

**Platforms**
- Win32, Linux, Solaris, and other UNIX

**Modes of operations**
- Online as a CORBA/EJB/RMI/COM Module
- Batch mode with a flat file input
  - For one-time runs

**Available for Oracle, other DB’s to follow**

System is delivered fully customized for the client’s database by ChoiceMaker staff
ChoiceMaker

Management
- Andrew Borthwick, President
  - Designed and implemented MEDD
  - NYU CS PhD 1999
  - Expert on maximum entropy modeling
- Arthur Goldberg, VP Strategy and Marketing
  - NYU CS Professor, co-director MSIS graduate program
  - Expert on network performance
  - Five years at IBM Research
- Staff includes three other Ph.D. computer scientists

Funding
- NSF Small Business Innovation Research Grant
- Investment from CCS, a $120M Japanese software firm

MEDD Features

Easy to Understand
- MEDD outputs a match probability, unlike other systems which output a "score"

Highly Customizable
- Powerful Java-based environment for creating custom comparison functions
- Advanced machine learning technology learns the human intuition for computing overall probability that a record-pair matches

Highly Accurate
- NYC DOH measured it as equivalent to two clerks working together

Questions

Arthur.Goldberg@choicemaker.com
Andrew.Borthwick@choicemaker.com
212 905-6031
ChoiceMaker Technologies, Inc.
41 East 11th Street, 11th Floor
New York, NY 10003