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.

CHOICEMAKER

MEDD

Maximum Entropy De-Duper

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

Prof. Arthur Goldberg, VP Marketing and Strategy
Dr. Andrew Borthwick, President

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

2

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

3

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

4

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 1998 clerks manually de-duplicated
  - 260,000 record pairs
  - 1,700 person-hours

MEDD De-Duplicates NYC Immunization Database

- Work in 1999-2000

Birth year	Records	Dupes removed
1996	203,389	25,553
1997	216,336	34,773
1998	208,315	47,830
1999	157,946	42,228
TOTAL	785,986	150,384

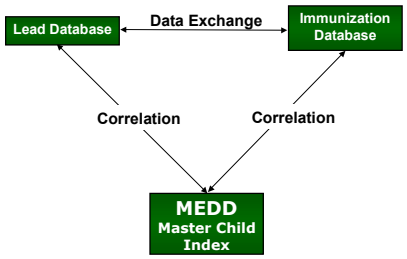
MEDD Links Two Databases

- 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

7

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



8

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

9

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)

10

Production Matching Basics

Input *Search record*

Blocking

- Find thousands of possible matches

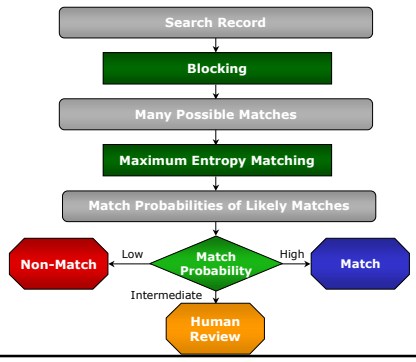
Match decision making

- 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*

11

Production Matching



12

Comparison Function Examples

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?

13

Comparison Function Examples

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?

14

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**

15

Customized with Java

Java-based Comparison Functions

- Simple first-name Soundex comparison function:

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

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

```
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();
}
```

16

Maximum Entropy Matching Math

- The probability a pair of records match

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

MatchProduct = product of weights of all comparison functions predicting **Match** for the pair

No-MatchProduct = product of weights of all comparison functions predicting **No-Match** for the pair

17

MEDD Decides Match
99.5% Confidence

Field Name	Record 1	Record 2	Match?	Weight
Last name	Smith	Smith	Match	1.153
First name	Emily	Emely	No-match	1.350
Soundex First name	EML	EML	Match	4.708
DOB	4/28/97	4/28/97	Match	1.138
Street	4528 3rd Ave	4528 3rd Ave	Match	4.342
City	Bronx	Bronx	Match	1.103
State	NY	NY		
Zip	10462	10462	Match	3.013
Phone	718-123-4567	718-123-6789	No-match	2.130
Med Rec Number	11856437503	11856437503	Match	6.587

Match product = 587.2
No-Match product = 2.9

$$\frac{587.2}{587.2 + 2.9} = 0.995$$

18

MEDD Decides No-match
97.9% Confidence

Field Name	Record 1	Record 2	Comparison	Weight
Last name	Lopez	Lopez	Match	1.153
First name Soundex First name	Girl	Susan	No data	
DOB	1/11/97	1/2/97	No-match	28.949
Street	987 Cornelia	456 Park	No-match	2.937
City	Brooklyn	Brooklyn	Match	1.103
State	NY	NY		
Zip	11211	11211	Match	3.013
Phone	718-123-4567	718-234-5678	No-match	2.130
Med Rec Number	1001002	567435		

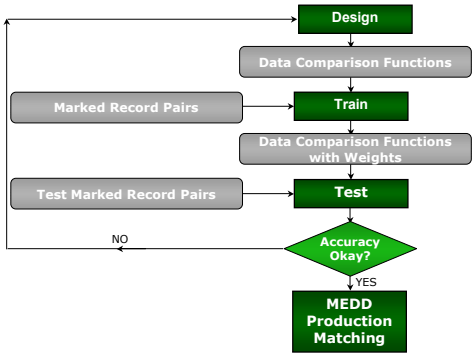
MatchProduct = 3.8

No-MatchProduct = 181.1

$$\frac{3.8}{181.1 + 3.8} = 0.021$$

19

MEDD Deployment

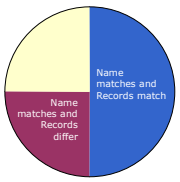


20

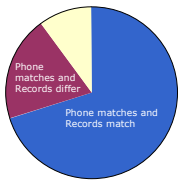
Principles of Maximum Entropy

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

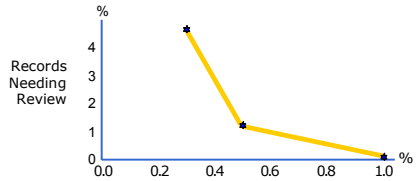
21

Demo

22

Matching with MEDD

Threshold Probabilities
Tradeoff Human Review Against Accuracy



Match Error Rate, NYC Immunization data

23

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

24

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

25

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

26

CHOICEMAKER

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
www.ChoiceMaker.com