Technical
White Paper

Probabilistic person matching

Scalable and accurate linking of healthcare data from multiple sources aggregated in the IBM Explorys Platform

Watson Health™

Contents

Introduction

Correctly matching records is necessary for accurate quality measures, financial reporting, and population management, but determining which healthcare records belong to each person is a challenging problem. Some data sources contain limited patient identifiers (e.g., laboratory feeds or insurance claims) and at other times demographic data elements conflict (e.g., asynchronous databases or basic human data entry errors).

Because multiple data sources are typically aggregated in the IBM® Explorys Platform, not all of which have reliable identifiers for use in linking records, the IBM Explorys Platform provides an accurate and scalable probabilistic person matching algorithm. The algorithm is configurable to match data from a variety of different sources to help meet the different needs and use cases encountered along with Watson Health™'s partners.

Missing data fields, different contact information, nicknames, changing last names, and/or simple typographical errors are obstacles a matching algorithm must overcome. For example, all but the last row in the table below might belong to the same person, so a holistic approach to collating records is necessary.

Name	Sex	Birth date	Address	MRN
Janice Doe	F	1987-04-02	123 Cedar, Cleveland OH 44106	A12345
Janice Doe	F	1978-04-02	123 Cedar, Cleveland OH 44106	A12345
Janice Doe	F	1987-04-02	123 Cedar, Cleveland OH 44106	
Jan Doe	F	1987-04-12		A12345
Jan Doe	F	1987-04-02	123 Cedar, Cleveland OH 44106	A12345
Janice Smith	F	1987-04-02	987 Main, Cleveland OH 44118	A12345
Janice Smith	F	1955-04-12	543 Euclid, Cleveland OH 44106	C98164

Note: The names and information that appear in the figures in this paper are used fictitiously for sample purposes only, and any resemblance to actual persons is entirely coincidental.

Probabilistic matching

A commonly encountered situation is that some records contain a unique identifier (e.g., MRN or EMPI) while other records don't have this identifier (e.g., laboratory feeds). In these situations, deterministic algorithms can have difficulties connecting records accurately.

The probabilistic algorithm makes use of identifiers, such as MRN and EMPI, when they are available, but can also incorporate many other data elements to compute an overall matching of records as described below.

The Fellegi-Sunter framework published in 1969 describes a mathematically optimal approach for record linkage and forms the backbone of the Watson Health probabilistic person matching algorithm. Adapting this framework to create a scalable solution and correctly modeling all relevant data elements (see "Features") are the key ingredients of the Watson Health solution for record matching.

Features

The record matching calculations are performed using the following information:

- First name, middle name/initial, last name
- Gender, birth year/month/day
- Street address, city, state, postal code
- Phone number, email address
- SSN, EMPI, MRN

Each feature is sent through a cleaning process before being used in the likelihood calculations. For example, SSN values are verified to be nine numeric digits that don't start with "666" while name values are parsed using more complex logic to handle prefixes and suffixes (like Dr. and Jr.), hyphenated names, and middle initials. Each individual feature has an associated model to best leverage the associated information. For example, first names are compared using an extensive database of nicknames, and last names have change for females at a statistically higher rate than for males.

Computational complexity

As shown in Figure 1, for N patient records, there are over N2 possible pair-wise comparisons that can be made. For a reasonable value for N of one million, this corresponds to one trillion comparisons! Watson Health uses a standard technique called blocking to reduce the computational complexity of the probabilistic person matching algorithm. Blocking partitions records into smaller subsets by only processing records with known similarities.

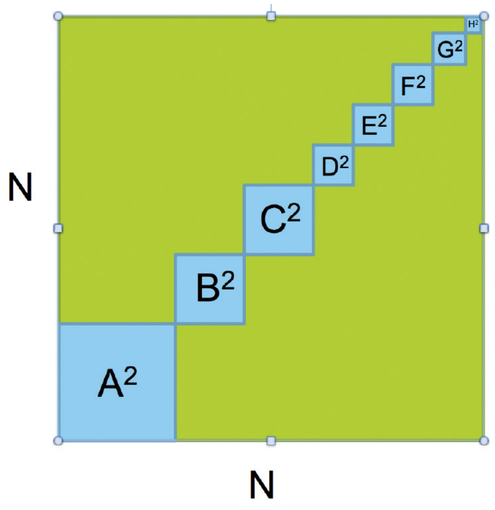

Figure 1: Visualization of run-time gains due to blocking. Green area represents complexity all pairwise comparisons, while sum of blue areas represent complexity of comparisons with appropriate blocking schemes

The Watson Health algorithm uses blocking schemes have been designed to compute pair-wise record comparisons with a reasonable chance of being a match. By analogy in the figure, this would allow for computational cost of the area in the blue boxes versus the entire green square.

One block might target patients whose records contain a typo in their last name, by making all pair-wise comparisons for records that have the exact same first name, birth date, and gender; another block might target patients with a typo in their birth date by filtering on exact first name, last name, and gender.

The Watson Health algorithm is scalable, fast, robust, and configurable. It has been designed to utilize Apache Hadoop1, which forms the foundation of the IBM Explorys Platform. This parallel computing framework helps enable the Watson Health algorithm to scale well to large populations. For a real dataset of 4.1M people with 17.1M demographic records, full run time is about 15 minutes.

Algorithm performance

On a test set of about 1M real demographic records, the type I error rate was estimated to be less than 0.1% and the type II error rate was estimated to be less than 1%.

Performance matrix	Actual match	Actual non-match
Declared match	30.37%	0.01%
Declared non-match	0.24%	69.38%

Extrapolated sensitivity	Extrapolated specificity
99.2%	99.9%

Algorithm performance will depend on the particular characteristics of the datasets being ingested, with the trade off between false positives (incorrect matches or type I errors) and false negatives (missed matches or type II errors) being customizable per customer requirements.

The Watson Health probabilistic person matching algorithm helps enable Watson Health customers to combine data from multiple sources in a scalable, configurable, and accurate way, and plays an important role in improving the value of the IBM Explorys Platform.

References

1. A Theory for Record Linkage, Ivan P. Fellegi and Alan B. Sunter, Journal of the American Statistical Association, Vol. 64, No. 328 (Dec., 1969), pp. 1183- 1210

2. Overview of Record Linkage and Current Research Directions, William E. Winkler, Statistical Research Division U.S. Census Bureau, Feb. 8, 2006, Statistics #2006-2

3. A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record-Systems, Carnegie Mellon University, Mauricio Sadinle and Stephen E. Fienberg, arXiv:1205.3217v2 [stat.AP] 6 Feb 2013

About IBM Watson Health

In April 2015, IBM launched IBM Watson Health and the Watson Health Cloud platform. The new unit will work with doctors, researchers and insurers to help them innovate by surfacing insights from the massive amount of personal health data being created and shared daily. The Watson Health Cloud can mask patient identities and allow for information to be shared and combined with a dynamic and constantly growing aggregated view of clinical, research and social health data.

For more information on IBM Watson Health, visit: ibm.com/watsonhealth.

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
June 2016

This document is current as of the initial date of publication and may be changed by IBM at any time. This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The information in this document is provided "as is" without any warranty, express or implied, including without any warranties of merchantability, fitness for a particular purpose and any warranty or condition of non-infringement.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Statement of Good Security Practices: IT system security involves protecting systems and information through prevention, detection and response to improper access from within and outside your enterprise. Improper access can result in information being altered, destroyed or misappropriated or can result in damage to or misuse of your systems, including to attack others.

No IT system or product should be considered completely secure and no single product or security measure can be completely effective in preventing improper access. IBM systems and products are designed to be part of a comprehensive security approach, which will necessarily involve additional operational procedures, and may require other systems, products or services to be most effective. IBM does not warrant that systems and products are immune from the malicious or illegal conduct of any party.

Note: The names and information that appear in the figures in this paper are used fictitiously for sample purposes only, and any resemblance to actual persons is entirely coincidental.