

What is matching?

- Definition: Finding identical or similar data items or data groups across two or more data collections. Various types of matching:
 - Schema-to-schema matching
 - Schema-to-ontology matching
 - Data item-to-data item matching
- Matching can signal identity or similarity (to form clusters)
- Issues:
 - Uncertainty/probability of match
 - Criticality of mismatch

What we cannot do today...

- We cannot automate in generalizable way either of:
 - Ontology / metadata alignment & matching procedures
 - Data value alignment & matching procedures
- Why? Because we don't know how to build generalizable matching criteria, and we don't know how to characterize in generalizable way the meaning, nature, and evolution of the data
- What are the problems:
 - Similarity/identity criteria are not absolute: they depend on the task and situation (how 'close' a similarity constitutes an acceptable match?, e.g., identical, close, somehow related...)
 - Matching depends on the nature of the data (what type of data is it?, e.g., text, number...) and requires the types of data to be precisely defined (is *cost* the same as *price*? Has it changed across tables?)

So where should we aim, overall?

1. Formal(?) languages in which to specify task/situation and nature of data at hand
2. Method(s) to derive/define/specify from this the (formally) appropriate matching criterion/ia
3. Work on building new kinds of matching functions, and combining matching functions to optimize their performance
4. Methods to evaluate systems, repeatedly, in order to measure progress
5. Methods to specify changes/evolution in data
6. Studies on consequences of error, for different kinds of data

Where is there promise for research?

- Improving the performance and generality of *individual* matching criteria/functions:
 - Simple, direct term matching: string matches
 - Text (definitions, documentation, etc.): NL extraction of terms/phrases, and matching of them
 - Ontology/metadata structure matching: tree/network similarity
 - Groups of entities, matched simultaneously: dispersal matches
 - Data values: alignment based on value types (simple type matching) and value distributions (e.g., using mutual information)
- Investigating methods of *combining* these criteria, using trainable combination functions
- Developing a standardized evaluation methodology to drive research:
 - Standard data collections, with gold-standard (human) alignments
 - Standard series of tests for the community
- Investigating methods for effective human interaction to drive/guide the matching

Matching: Nature of the Problem

- Kinds of matching
 - Property-level: domain-value, content (feature-based), compound structure, context aware
 - Entity-, Object-level
- Outputs: Boolean, best-match, probability, similarity, top-k
- Procedures & operators for matching
- Task, application needs
 - Going between why are you matching <-> How should you be matching
- Uses of matching in II
 - Collation
 - Linking
 - Consolidation, duplicate detection
 - Ranking
 - Correction, Update

More Nature

- For what purpose
 - Further human attention
 - Annotate, connect records
 - Coalesce info

Our Threads

- Importance of task application, application context
 - selecting appropriate methods, parameters
 - understanding importance, consequences, criticality
- Moving from one-off, domain-specific matching, *to* Reusable, multi-use, *to* Generic, wide-use
- Using human attention effectively
 - re-use
 - directing use
- Granularity of match
- Moving from structured, textual to unstructured, multi-modal

More Threads

- Metrics – what to measure
 - Cost, speed, quality, completeness, scale, usability
- Light v. tight (best-effort v. precise)

Non-thread: We should have talked more about provenance of matching

Hard Problem, Approaches

Entity Purging: Removing all information related to an entity from a collection of information systems

Research Approaches

Contests

Intensive summer working groups

Challenge Problems

- Ad hoc integration: Impending hurricane scenario: risk assessment, response planning
 - What kinds of matching? Geospatial, imagery, text, ...
- Personal Health Record (PHR): Collect and organize health info from all providers
 - Each patient has a different set of sources
 - Different collections for different uses

Milestones

2 years

GiigleTM: Generic, lightweight information integration

Transferring human (and machine) work (e.g., mark-up, annotation)
across incremental revisions

- revisions of base info
- of overlaid info

Reuse frameworks for matching: Not going back to zero

5 years

Harnessing community contributions w/appropriate quality & trust

Integration center?

Learnable domain-specific, task-specific matchers

9.6 ± 0.7 years with 94% confidence

Probabilistic

Multi-modal

What can we do soon?

- Create an evaluation resource and community
 - Model on Speech, TREC, DUC, etc., in NL community
 - Gather standard sets of data collections, with their human mappings
 - Arrange friendly periodic competitions
 - Use these to explore specific community-wide problems
 - Implement and distribute one or more packages of fairly standard general/approximate matching functions (term matches, etc.)
 - This allows researchers to ramp up quickly
- Specify a set of more complex matching challenge problems:
 - Explore data-to-data matching routines
 - Explore schema-to-schema/ontology routines
 - Explore approximate matching / collation (e.g., in the case of data updates, evolution/change, or when things are known to not be identical)
 - Purging: remove a given entity from everywhere in a data collection

Don't overpromise

- Fully-automated high-accuracy matching is impossible (for the near and middle term)
- Handling data evolution and schema change is impossible for a long time