APPENDIX I

DEFINITIONS OF TERMS USED IN THIS THESIS

Active learning

A type of classifier that builds a classification model based on the similarity values in comparison vectors and optionally the relationships between pairs or groups of records. In order to improve its classification model, an active learning classifier asks for manual feedback on the match status of selected candidate record pairs (that are difficult to classify). An iterative process of manual feedback and building an improved classification model is carried out until a certain stopping criterion has been achieved.

Approximate matching

The process by which a candidate compares record pairs using a set of comparison functions that allows for approximate (not exact) similarities.

Attribute

A column in a database table, file, or spreadsheet, that contains a well defined type of data, such as strings, numbers, dates, times and so on.

Attribute value

A value stored in a specific attribute and a specific record in a database, file or spreadsheet.

Candidate record pair

It is a record pair, which is formed from the records that are inserted into the same block (or cluster or window) by an indexing technique. All candidate record pairs are compared in detail in the comparison step of the data matching process using various comparison functions.
Classification model

A model, that determines the classification of candidate record pairs into matches, non-matches and optionally potential matches. A classification model is generated by a classifier algorithm.

Classifier

A type of algorithm that builds a classification model based on a supervised learning, unsupervised learning or active learning approach.

Cleaning

The process of removing unwanted characters and tokens (alphanumerical words) from the attribute values in the input database(s) during the data pre-processing step of the data matching process.

Clustering

A type of algorithm that groups similar data objects (records in the case of data matching) together, according to the similarities calculated between these objects. In data matching and deduplication, the similarities between records are captured in the comparison vector for each candidate record pair. The objectives of clustering in data matching and deduplication are to (1) have each generated cluster to correspond to one entity, and (2) each entity stored in the database(s) that are matched or deduplicated is assumed to be represented by one cluster only.

Comparison function

A function that has two attributes values as input (which values can be strings, numbers, dates, times or more complex objects) and that calculates a similarity between these two values. The comparison can either be exact or allow for approximate similarity. An exact comparison function generally returns a similarity value of 0 if the two attribute values are different from each other, or a similarity value of 1 if they are the same. An approximate comparison function generally returns a normalized numerical similarity value between 0
and 1 that indicates the similarity between the two attribute values, with a larger similarity value indicating a higher similarity between the two attribute values. If only exact comparison functions are used when attribute values are compared, then this process corresponds to exact matching, while when approximate comparison functions are used, the process corresponds to approximate matching. A popular class of approximate comparison functions is approximate string comparison functions.

Comparison vector

The vector of similarity values generated for a candidate record pair when one or more attributes of the pair are compared using comparison functions that are appropriate to the content of the attributes. If n comparison functions are used then the resulting comparison vector will contain n similarity values.

Data linkage

The name used by statisticians, and health and biomedical researchers and practitioners, for the process of data matching.

Data matching

It is the process of comparing records from two or more databases with the objective of identifying pairs or groups of records, which refer to the same entity. These pairs or groups of records are known as matches.

Data pre-processing

The process of cleaning, standardizing, and segmenting, the attribute values stored in the input database(s), to be matched or deduplicated, with the general aim to improve data quality, and more specifically, to improve the outcomes of the data matching or deduplication process.

Deduplication

The process of duplicate detection followed by a process which, for each entity in a database, either merges the identified duplicate records into one combined record, or removes some records from the database until it contains only single
record for each entity.

Duplicates

The presence in a single database of multiple records, that refers to the same entity.

Duplicate detection

It is the process of comparing records from a single database with the objective of identifying pairs or groups of records that refer to the same entity. These pairs or groups of records are known as duplicates.

Entity

A real-world subject or object, such as an individual person, business, publication or consumer product, that has a unique identity and that can be distinguished from any other entity.

Entity identifier

A number, code or string that uniquely identifies a single entity within the database(s) that are matched or deduplicated.

Entity resolution

The process of comparing records from one or more databases with the objective of identifying pairs or groups of records that refer to the same entity, to classify these pairs or groups as matches (and pairs or groups of records that do not refer to the same entity as non-matches), and to merge all records that refer to the same entity into a new combined record. The result of an entity resolution process is a set of combined records that each corresponds to one entity, and a single combined record represents each of the entities stored in the database(s) that were matched only. An Entity resolution applied on a single database is also known as deduplication.
**Exact match**

The status of a candidate record pair that has been compared using either an exact or an approximate matching process, and where all attribute values that have been compared are the same in both records of the pair.

**Exact matching**

The process by which, candidate record pairs are compared using a set of comparison functions that only permit exact similarities.

**Indexing**

It is the process of splitting a database into smaller blocks or clusters, or sorting out? A database, with the objective of reducing the number of record pair comparisons that are conducted. Records that have the same blocking key value are inserted into the same block or cluster, or they are sorted close to each other if they have the same or a very similar sorting key value. Candidate record pairs are formed from all records that are in the same block or cluster, or that appear in the same window. The traditional approach used for indexing has been blocking.

**Match**

It is a pair or group of records classified as referring to the same entity.

**Merge/purge**

The name used by database and data warehousing researchers and practitioners for the process of data matching and deduplication.

**Non-match**

It is a pair or group of records that is classified as referring to different entities.
Phonetic encoding

It is a type of algorithm that converts a string (generally assumed to correspond to a name) into a code that represents the pronunciation of that string. Popular phonetic encoding algorithms include Soundex, NYSIIS, ONCA, Phonex, Phonix and Double-Metaphone.

Privacy-preserving data matching

Also known as privacy-preserving record linkage, this is the process of matching databases from different organizations such that none of the database owners has to reveal any of their private or confidential data, and at the end of the matching process only limited information, such as the number of records that have been classified as matches, or only their record identifiers, is being revealed to the database owners.

Probabilistic record linkage

A statistical classifier approach to data matching published by Fellegi and Sunter in 1969. This approach calculates match weights and non-match weights based on error probabilities and frequency distributions of attribute values in the input databases. Candidate record pairs are classified based on their weight vectors into either matches, non-matches, or potential matches, using a threshold-based and pair-wise classification approach.

Record

It is a row in a database table, file, or spreadsheet that contains values in a set of attributes. It is assumed that each record represents one entity, but that an entity can be represented by more than one record in a database, file or spreadsheet.

Record identifier

It is a number, code, or string, which uniquely identifies a single record in a database. A record identifier is different from an entity identifier.
Record pair

Two records, for the process of data matching one record from each of the two input databases that are matched, while for the deduplication of one database both records are sourced from the single input database.

String comparison function

A type of comparison function that takes as input two strings and that returns an exact similarity value (exact matching) or an approximate similarity value (approximate matching) calculated for the two input strings.

Supervised learning

A type of classifier algorithm that builds a classification model based on the similarity values in comparison vectors and optionally the relationships between pairs or groups of records. A supervised classification model is built based on training data that are in the form of pairs or groups of records where their match status (true match and true non-match) is known.

True match

It is a record pair that is classified as a match, where in both records in the pair correspond to the same entity. In the context of classification, a true match is also known as a true positive.

True non-match

It is a record pair that is classified as a non-match, where in the two records in the pair correspond to two different entities. In the context of classification, a true non-match is also known as a true negative.
Weight vector

It is a vector containing numerical values for a candidate record pair. Weight vectors are used by probabilistic record linkage and related classifier approaches to decide the match status of candidate record pairs. The values in the weight vector of a candidate record pair are calculated by combining for each compared attribute the match weight (if attribute values are the same or similar) or non-match weight (if attribute values are different) for that attribute with the similarity value for that attribute taken from the pair’s comparison vector.

Unsupervised learning

A type of classifier algorithm that builds a classification model based on the similarity values in comparison vectors and optionally the relationships between pairs or groups of records. The classification model is built without knowing the true match status of these pairs or groups of records. Clustering is a popular type of unsupervised learning algorithm.