## Data Intensive Systems Literature

### Mining Massive 3.5

The jaccard similarity is the measure how close sets are.

**Jaccard distance:** is the distance between sets (1-similarity) which is the function d(x,y)

**Distance measure conditions:**

* d(x,y) >= 0
* d(x,y) == 0 when x = y
* d(x, y) ≤ d(x, z) + d(z, y) when x,y,z in triangle

**Euclidean Distance**: most common way for distance measure

N-dimensional Euclidian space is one where points are vectors of n real numbers. The conventional distance measure aka the L\_2-norm is (because r=2 usually making it ½ aka root):

*Explaination: square the distance in each dimension, sum the distnace, take the positive square root*

**Manhattan distance**: the magnitude sum of the differences in each dimension (travel distance by using grid lines)

**Cosine distance:** For Dimension spaces where points are vectors with integer or bool components, the cosine distance is the angle that the vectors to those points make (0-180).

E.G. two points: x = [1,2,-1] and y = [2,1,1].

* **Dot product x.y** is 1 \* 2 + 2 \* 1 + (-1) \* 1 = 3
* L\_2 norm x is = and y is =
* Cosine distance x.y is 3/() = 0.5 and is 60 degrees

**Edit distance** is for string points where the distance is the smallest number of insertions and deletions of single characters to convert x to y. Each step is one point

**LCS edit distance** is calculated by only deleting positions from both x and y (can only be calculated when the length of x plus the length of y minus twice the length of their LCS)

E.G. x = abcde and y = acfdeg and LCS = acde

* Normal: remove b + add f + add g = 3 steps
* LCS: 5 + 6 − 2 × 4 = 3

**Hamming distance** is the number of components that 2 vectors differ. (10101 and 11110 = 3)

### Record detection

analysis of the literature on duplicate record detection using similarity metrics that are commonly used to detect similar field entries, and we present an extensive set of duplicate detection algorithms that can detect approximately duplicate records in a database + improving the efficiency and scalability of approximate duplicate detection algorithms.

Data heterogeneity: an error when intergrating data from different sources

* Structural: different way of typing column data (street vs adrr)
* **Lexical: different representations to refer to the same object (St. vs Street)**

Data preparation steps:

1. **Data parsing** locates, identifies, and isolates individual data elements in the source files.
2. **Data transformation** refers to simple conversions that can be applied to the data in order for them to conform to the data types of their corresponding domains.
3. **Data standardization** refers to the process of standardizing the information represented in certain fields to a specific content format.

#### Field matching techniques

##### Character-based similarity metrics

Fix typo errors

* **Edit distance** is for string points where the distance is the smallest number of insertions and deletions of single characters to convert x to y. Each step is one point
* **Affine gap distance**: Fix truncated/shortened strings (such as names) using open and extend gaps
* **Smith-waterman distance:** Fix truncated strings (like prof \*name\* or \*name\* prof) using placement distance
* **Jaro distance:** Fix last and first name strings using common characters and the Jaro comparison value
* **Q-gram distance:** when two strings 1 and 2 are similar, they share a large number of q-grams in common.

##### Token-based similarity metrics

Fix typo errors by ignoring the left-right placement (also use right-left)

* **Atomic strings:** a sequence of alphanumeric characters delimited by punctuation characters (for prefixes)
* **WHIRL:** a information retrieval tool the cosine similarity combined with the tf.idf weighting scheme to compute the similarity of two fields
* **Q-grams:** extended the WHIRL system to handle spelling errors by using q-grams, instead of words, as tokens.

##### Phonetic similarity metrics

Fix Spoken/tone similar words typos (cajun and kageonne sounds the same)

* **Soundex:** the assignment of identical code digits to phonetically/tone similar groups of consonants and is used mainly to match surnames.
* **NYSIIS:** retains information about the position of vowels in the encoded word by converting most vowels to phonetically similar letters/code
* **ONCA:** uses a British version of the NYSIIS method of compression and Soundex-ed in the usual way
* **Metaphone:** using 16 consonant sounds that can describe a large number of sounds used in many English and non-English words
* **Double metaphone:** allowing multiple encodings for names that have various possible pronunciations

##### Numeric similarity metrics

None yet

#### Detecting duplicate records

### Map Reduce

**MapReduce** is a programming model and an associated implementation for processing and generating large data sets. Users specify a **map** function that processes a key/value pair to generate a set of intermediate key/value pairs, and a **reduce** function that merges all intermediate values associated with the same intermediate key.

## Data Stream

*Data arrives in a stream or streams, and if it is not processed immediately or stored, then it is lost forever.*

To summarize the stream, one approach is to filter the desirable elements and estimate the total elements given a part of the stream. Another is to look at only a fixed-length “window” consisting of the last n elements for some (typically large) n. We then query the window as if it were a relation in a database. If there are many streams and/or n is large, we may not be able to store the entire window for every stream, so we need to summarize even the windows.

Query types:

* Standing query: producing a query with a standard format/calculator/algorithm (max or average)
* Ad-hoc: adding a window with a filter to get the best information (users with most active website time)

A diagram of a data stream management system

Description automatically generated