Data Matching: Basic Concepts

Applies to:

Automated Matching (De-Duplication, Consolidation) of Data Records.

SAP Data Services, Data Quality Management.

For more information, visit the Data Services and Data Quality Space.

Summary

Each implementation of Data Matching inevitably raises the question of an adequate configuration of the matching engine(s) that support the matching process. Whether we only intend to run rare searches for duplicate records in a single data base, or we are expected to implement a fully autonomous complex architecture to match and consolidate the various types of data (companies, contact persons, transactions, etc.) arriving from a multitude of disparate data sources – we will need to apply a number of basic concepts that go with any Data Matching functionality.

In this paper we provide a brief overview of those basic concepts in a relatively simple and visual manner.

Author: Sergey LUKYANCHIKOV

Company: SAP France S.A.

Created on: November 23, 2012

Author Bio

Sergey LUKYANCHIKOV, SAP France S.A, is an Industry Analytics Principal at SAP Performance & Insight Optimization (SAP PIO), specialist in Energy, Utilities and Services.

Table of Contents

Fuzzy vs. Exact Matching

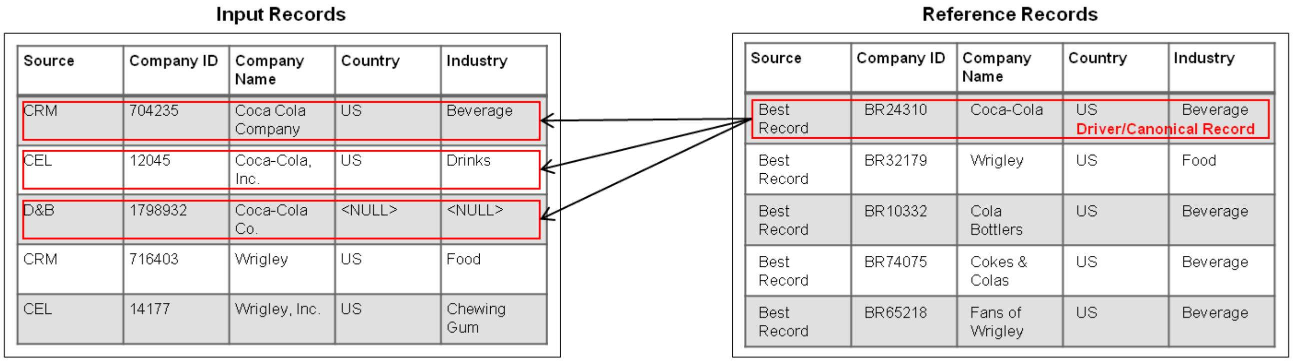
Matching of data records can be based on a number of exact criteria: for example, if the first name is “John” and the last name is “Bull”, then match all such records into the same group. However, should there be a record with a slight misspelling in the first name – “Jhon” instead of “John” – the exact matching mechanism will ignore such a record, while the misspelled record should belong to the same group as the other records with the first name spelled correctly. To overcome such a limitation, matching mechanisms based on fuzzy logic are used. The essence of fuzzy logic is that instead of the two outcomes that exist in the binary logic – “match” and “no match” – it operates probabilities of a match that range from 100% to 0%.

Matching Reference Data

Most of the practical implementations of fuzzy matching mechanisms presume that there is a given set of data records (input data) in which for every record a probability of a match with a record in another set of data records (reference data) must be evaluated.

Role of Driver/Canonical Records in Data Matching

A record in the reference data set against which the probability of a match for a selected record from the input data set is calculated is called the **driver record** or **canonical record**. The approach to the definition of the driver record depends on the functionality of a concrete matching engine. In general, those approaches can be split into two groups: those based on **direct referencing** (a single driver record is identified in the reference data against which the matches of several similar records from the input data are evaluated) –



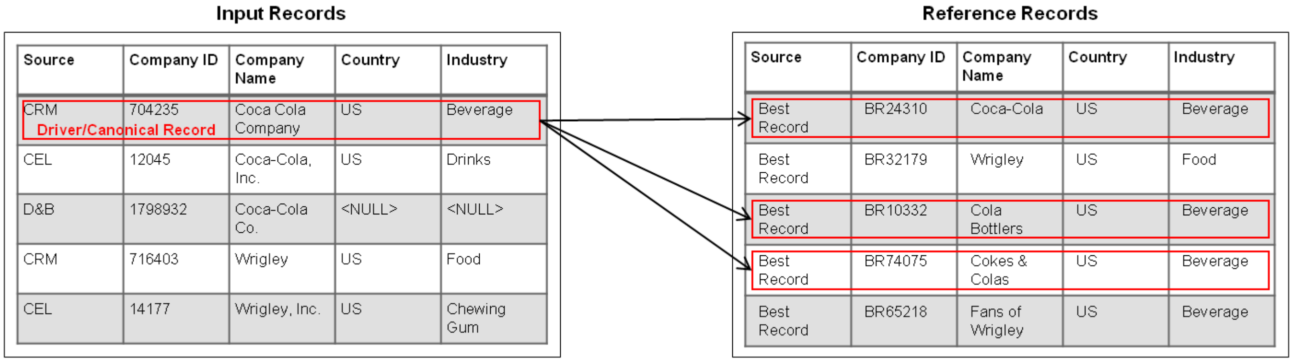
Input Records

Source	Company ID	Company Name	Country	Industry
CRM	704235	Coca Cola Company	US	Beverage
CEL	12045	Coca-Cola, Inc.	US	Drinks
D&B	1798932	Coca-Cola Co.	<NULL>	<NULL>
CRM	716403	Wrigley	US	Food
CEL	14177	Wrigley, Inc.	US	Chewing Gum

Reference Records

Source	Company ID	Company Name	Country	Industry
Best Record	BR24310	Coca-Cola	US	Beverage
Best Record	BR32179	Wrigley	US	Food
Best Record	BR10332	Cola Bottlers	US	Beverage
Best Record	BR74075	Cokes & Colas	US	Beverage
Best Record	BR65218	Fans of Wrigley	US	Beverage

– and those based on **indirect** or **reversed referencing** (for a record from the input data, the matches of several similar records from the reference data are evaluated):



Input Records

Source	Company ID	Company Name	Country	Industry
CRM	704235	Coca Cola Company	US	Beverage
CEL	12045	Coca-Cola, Inc.	US	Drinks
D&B	1798932	Coca-Cola Co.	<NULL>	<NULL>
CRM	716403	Wrigley	US	Food
CEL	14177	Wrigley, Inc.	US	Chewing Gum

Reference Records

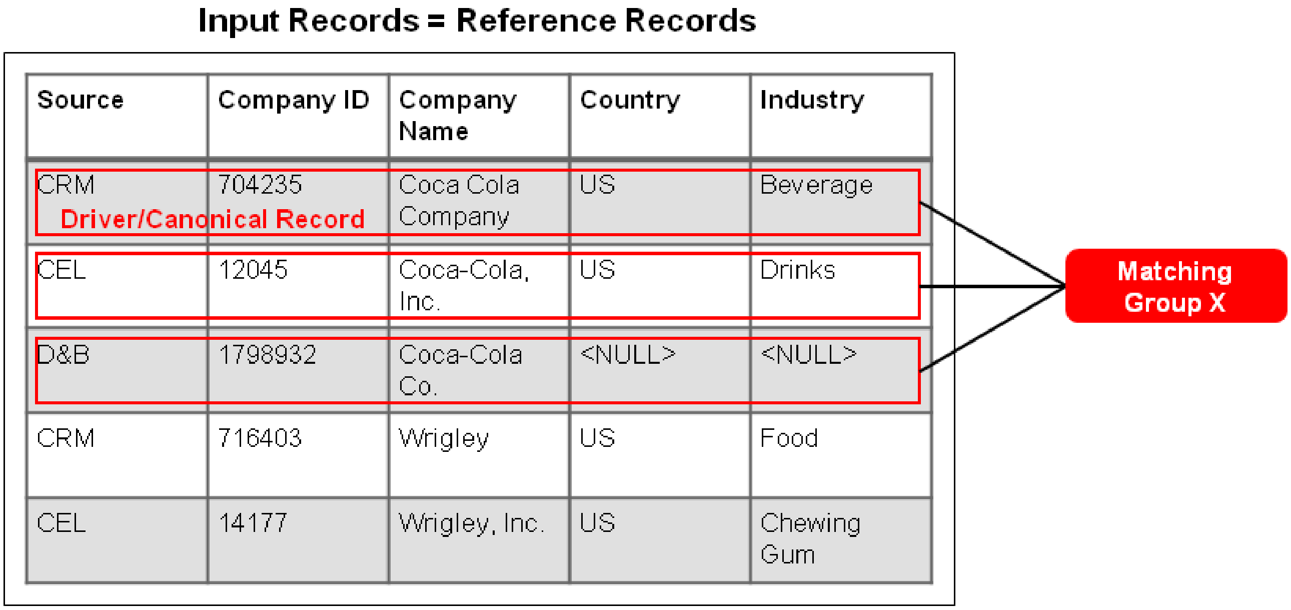
Source	Company ID	Company Name	Country	Industry
Best Record	BR24310	Coca-Cola	US	Beverage
Best Record	BR32179	Wrigley	US	Food
Best Record	BR10332	Cola Bottlers	US	Beverage
Best Record	BR74075	Cokes & Colas	US	Beverage
Best Record	BR65218	Fans of Wrigley	US	Beverage

Direct referencing reposes on the assumption that the reference data is carefully de-duplicated and that the records in the reference data with high degree of probability are superior in terms of completeness and data actuality relative to any of the similar records in the input data.

Indirect or reversed referencing admits that the reference data may contain duplicates (e.g., one of the non-de-duplicated sources, like CRM or ERP, is taken as a reference because of the access speed or lookup performance reasons) and leads to the appearance on the list of resulting matches of multiple matches for one input record. If processed regularly via a best record consolidation engine and returned back to matching and matched again and again with the newly received input data, the results of matching against such an “imperfect” reference data may still provide better consolidated result (less fragmented) than in the case of the absence of the reference data (i.e., in the case of the input data used as reference).

Input Data as Reference Data

The most straightforward approach to the implementation of reference data is the use of input data as reference data. In this case, a matching mechanism identifies driver records among the input records and performs match evaluation of the other records against those driver records:



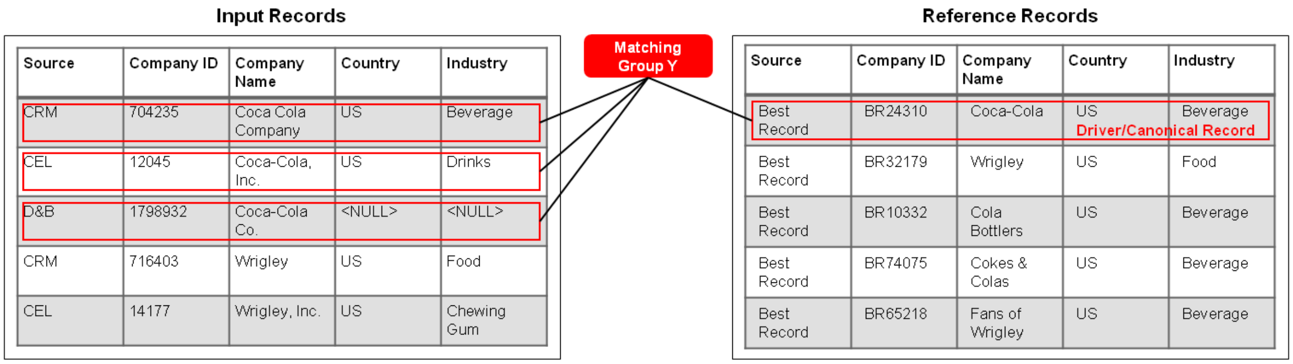
Source	Company ID	Company Name	Country	Industry
CRM	704235	Coca Cola Company	US	Beverage
CEL	12045	Coca-Cola, Inc.	US	Drinks
D&B	1798932	Coca-Cola Co.	<NULL>	<NULL>
CRM	716403	Wrigley	US	Food
CEL	14177	Wrigley, Inc.	US	Chewing Gum

Technically, the input record table is matched against its temporary copy, and the temporary copy acts as reference data for the original input record table. The two referencing modes (direct and indirect) in this case would have no distinctions because both would deliver two identical sets of “driver-subordinate” pairs of records all of them identified in the original input record set and all of them subject to a further calculation of the respective similarity scores.

External Data as Reference Data

A more elaborate approach is: to use with each new portion of input data, as reference data, the database of de-duplicated (ideally, via best record consolidation) results of the previous matching runs. De-duplication via best record consolidation would deliver unique (within the limits of a specific matching group) and more complete/actualized records (compared to the previous disparate input records).

Or alternatively, an authoritative source of information (like the corporate CRM system) that is likely to contain validated and most up-to-date records (although frequently with a fair share of duplicates) can be used as “compromise” reference data.



Input Records

Source	Company ID	Company Name	Country	Industry
CRM	704235	Coca Cola Company	US	Beverage
CEL	12045	Coca-Cola, Inc.	US	Drinks
D&B	1798932	Coca-Cola Co.	<NULL>	<NULL>
CRM	716403	Wrigley	US	Food
CEL	14177	Wrigley, Inc.	US	Chewing Gum

Reference Records

Source	Company ID	Company Name	Country	Industry
Best Record	BR24310	Coca-Cola	US	Beverage
Best Record	BR32179	Wrigley	US	Food
Best Record	BR10332	Cola Bottlers	US	Beverage
Best Record	BR74075	Cokes & Colas	US	Beverage
Best Record	BR65218	Fans of Wrigley	US	Beverage

The huge advantage of the use of external reference data is the quicker "convergence" of the completeness and actuality of the disparate input records to a stable best record state – the records in a newly received portion of input data would more and more likely "find" their respective driver records in the more and more complete and actual reference database, and would quicker "profit" from the superior quality of the drivers (i.e. would either become discarded as non-contributors to the further quality improvement of the consolidated best record, or would contribute the missing pieces of information to the best record via a best record consolidation run).

Similarity Score Calculation

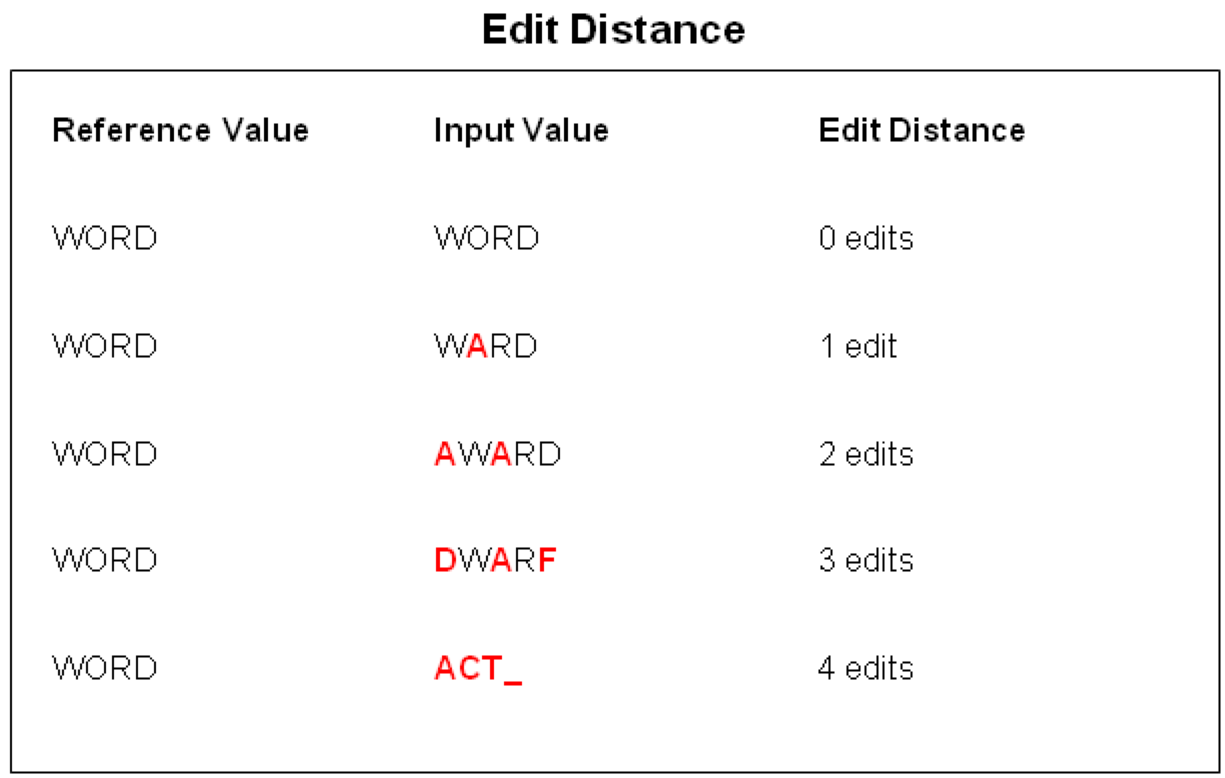
The evaluation of the match probability, the primary objective of any fuzzy matching mechanism, must end up in a concrete numeric indicator – a probability of a match expressed either as a percentage (100% to 0%), or as a value in the range of [1; 0], or as any other numeric value associated with a numeric interval (continuous or in some cases discrete – e.g. due to rounding up of the indicator values to the closest integer or decimal values in the interval) that represents the two "pure" cases ("match" and "no match") as well as the "in-between" evaluations of the match probability.

Role of Similarity Score in Data Matching

The numeric indicator mentioned above is called the **similarity score** (sometimes also the match score) – it expresses the degree of similarity between a record from the input data and a driver record (or a driver record from the input data and a reference record – if reversed referencing is used). The similarity score calculation can be preceded or not by an execution of a number of auxiliary business rules (e.g., replace several synonym words with a single synonym, replace abbreviations by full text labels, remove any "nonsense" words, etc.) to help the matching mechanism to calculate a "more realistic" similarity score. The similarity score calculation can be based on one or a combination of methods. Below we explain the two most common methods – **edit distance** and **phonetic equivalence** (sometimes also referred to as "soundex").

Similarity Score Based on Edit Distance

The edit distance between two texts is the number of characters it takes to change in one text to obtain the other:



Edit Distance

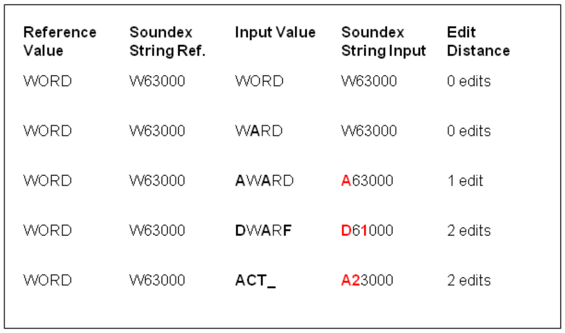
Reference Value	Input Value	Edit Distance
WORD	WORD	0 edits
WORD	WARD	1 edit
WORD	AWARD	2 edits
WORD	DWARF	3 edits
WORD	ACT_	4 edits

In some of the commercially available matching engines it is this method that serves as a basis for evaluating the probability of a match between an input text and a reference text. Of course, in the commercial matching engines the input and the reference texts are both subdivided into smaller portions, and the similarity score calculation proceeds via a comparison of those smaller portions. Nevertheless, to compare two smaller portions of the original texts, the edit distance is taken as a primary measure of difference.

Similarity Score Based on Phonetic Equivalence

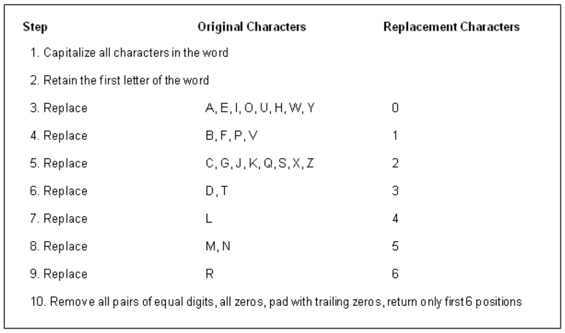
Two texts are phonetically equivalent if their phonetic values (i.e., the soundex strings) are the same or the difference between them is within the specified tolerance limits.

Phonetic/Soundex Equivalence (Measured via Edit Distance)



Reference Value	Soundex String Ref.	Input Value	Soundex String Input	Edit Distance
WORD	W63000	WORD	W63000	0 edits
WORD	W63000	WARD	W63000	0 edits
WORD	W63000	AWARD	A63000	1 edit
WORD	W63000	DWARF	D61000	2 edits
WORD	W63000	ACT_	A23000	2 edits

Phonetic/Soundex Algorithm (English)*



Step	Original Characters	Replacement Characters
1. Capitalize all characters in the word		
2. Retain the first letter of the word		
3. Replace	A, E, I, O, U, H, W, Y	0
4. Replace	B, F, P, V	1
5. Replace	C, G, J, K, Q, S, X, Z	2
6. Replace	D, T	3
7. Replace	L	4
8. Replace	M, N	5
9. Replace	R	6
10. Remove all pairs of equal digits, all zeros, pad with trailing zeros, return only first 6 positions		

* A generic version, could differ depending on concrete implementation

Take the example of WORD and WARD. If the edit distance is measured directly among these two texts, they are 1 edit away one from another. But if we compare their respective soundex strings, they result to be equivalent (see the above diagram). Therefore, the phonetic equivalence mechanism while tolerating a certain approximation in the course of the comparison of two texts, at the same time opens additional possibilities for an error-tolerant similarity score calculation – occasional non-critical misspellings will be worked around.

Similarity Score Based on a Combination of Methods

In a number of implementations of matching mechanisms there exists a possibility to combine (via formulas) several methods to calculate a similarity score.

Use of Match Thresholds

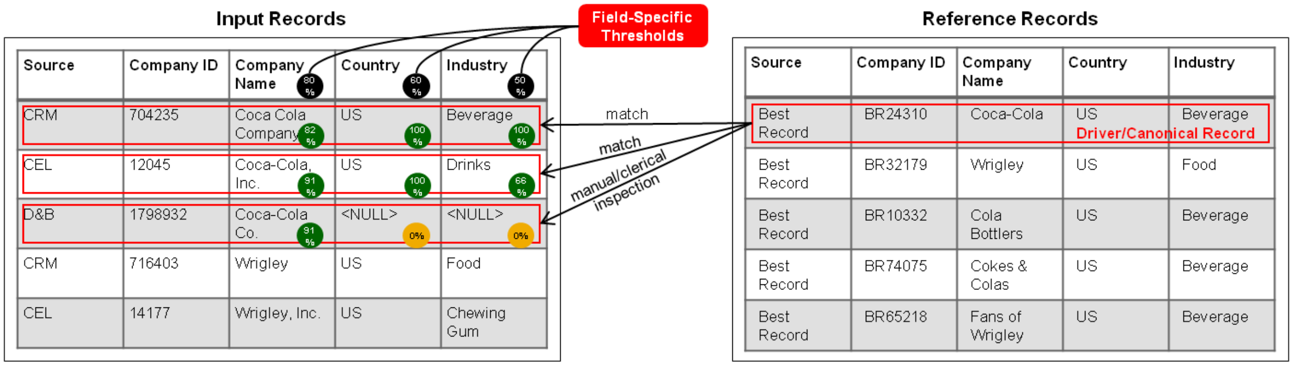
Once the similarity scores are calculated for all of the input records, the matching mechanism must discriminate (based on the similarity scores) between the "match" and "no match" cases.

Role of Match Thresholds in Data Matching

The common approach to discriminate between the "match" and "no match" cases is the use of **match thresholds**: the constants defined either per matching field or per whole matching scenario and expressing the minimum similarity score value that is required for a "match" case. Some implementations of matching mechanisms provide more than just a single "match/no match" threshold, but allow to define two separate thresholds: one for "match" cases (similarity scores that are equal or greater), one for "no match" cases (similarity scores that are equal or less), assuming that the cases that fall between the two thresholds are verified and resolved manually by an operator.

Sequential Application to Matching Fields of Field-Specific Match Thresholds

One of the ways to apply match thresholds is to sequentially (one by one) verify that every matching field similarity score value on an input record passes its respective "field-specific" match threshold. If a least one matching field similarity score value fails to pass its threshold, the input record is considered a "no match" (or is quarantined for a manual resolution by an operator).



Input Records

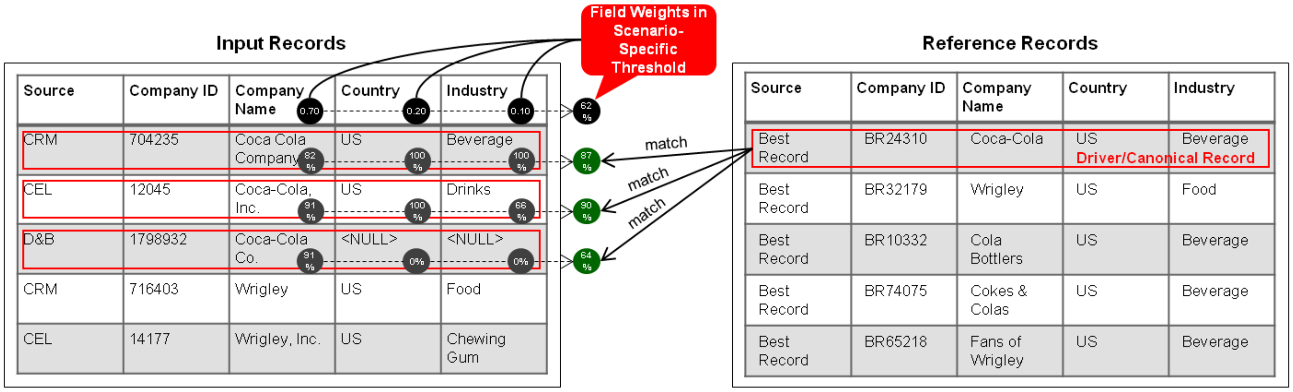
Source	Company ID	Company Name	Country	Industry
CRM	704235	Coca Cola Company	US	Beverage
CEL	12045	Coca-Cola, Inc.	US	Drinks
D&B	1798932	Coca-Cola Co.	<NULL>	<NULL>
CRM	716403	Wrigley	US	Food
CEL	14177	Wrigley, Inc.	US	Chewing Gum

Reference Records

Source	Company ID	Company Name	Country	Industry
Best Record	BR24310	Coca-Cola	US	Beverage
Best Record	BR32179	Wrigley	US	Food
Best Record	BR10332	Cola Bottlers	US	Beverage
Best Record	BR74075	Cokes & Colas	US	Beverage
Best Record	BR65218	Fans of Wrigley	US	Beverage

Simultaneous Application to Matching Fields of a Scenario-Specific Match Threshold

Another way to apply match thresholds is to simultaneously (all matching fields at once) verify that either the smallest, or the simple average, or the weighted average similarity score of the matching fields across the input records, passes the scenario-specific threshold.



Input Records

Source	Company ID	Company Name	Country	Industry
CRM	704235	Coca Cola Company	US	Beverage
CEL	12045	Coca-Cola, Inc.	US	Drinks
D&B	1798932	Coca-Cola Co.	<NULL>	<NULL>
CRM	716403	Wrigley	US	Food
CEL	14177	Wrigley, Inc.	US	Chewing Gum

Reference Records

Source	Company ID	Company Name	Country	Industry
Best Record	BR24310	Coca-Cola	US	Beverage
Best Record	BR32179	Wrigley	US	Food
Best Record	BR10332	Cola Bottlers	US	Beverage
Best Record	BR74075	Cokes & Colas	US	Beverage
Best Record	BR65218	Fans of Wrigley	US	Beverage

Application of Match Thresholds Based on a Combination of Methods

Depending on the implementation of a matching mechanism, various combinations of match threshold application methods can be possible. For example, some part of the matching fields may undergo a sequential verification of their field-specific thresholds, while the resting matching fields may have their similarity scores compared with the scenario-specific threshold defined for the given matching scenario.

Matching Results Evaluation

In the end of any matching process it is important that the matching results (“match” and “no match” cases identified by the matching mechanism) are evaluated – verified by an operator, supplied with matching quality indicators and statistics.

Positive and Negative, True and False Matches

In a more strict terminology, instead of “match” and “no match” cases, the following outcomes are identified:

- **True positive match:** the input record and the driver record had to be matched and they were matched
- **True negative match:** the input record and the driver record did not need to be matched and they were not matched
- **False positive match:** the input record and the driver record did not need to be matched but they were matched
- **False negative match:** the input record and the driver record had to be matched but they were not matched

Manual/Clerical Evaluation

Manual (also referred to as **“clerical”**) **evaluation** of matching results is based on a visual control of the matches by an operator.

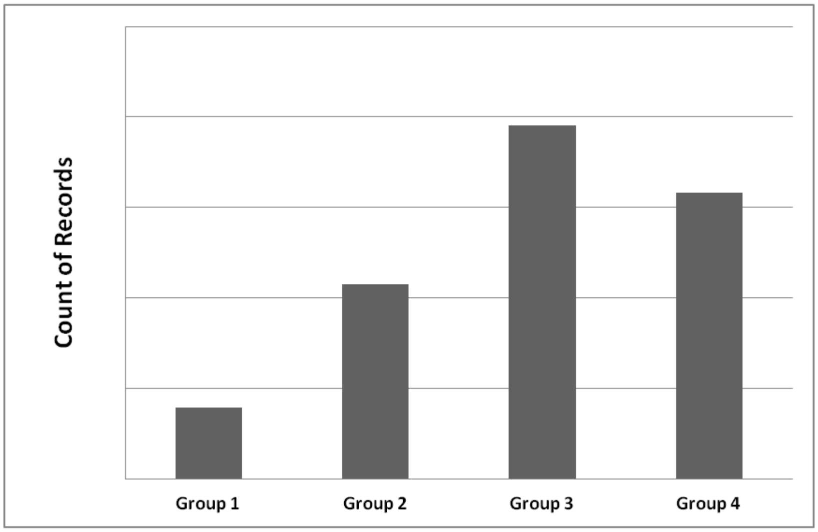
Match Rate Calculation

The most frequently used indicator to assess the overall quality of verified matching results is the **match rate**. The match rate calculation may differ depending on the particular approaches to the calculation of “match” and/or “no match” counts, but commonly it follows the following formula:

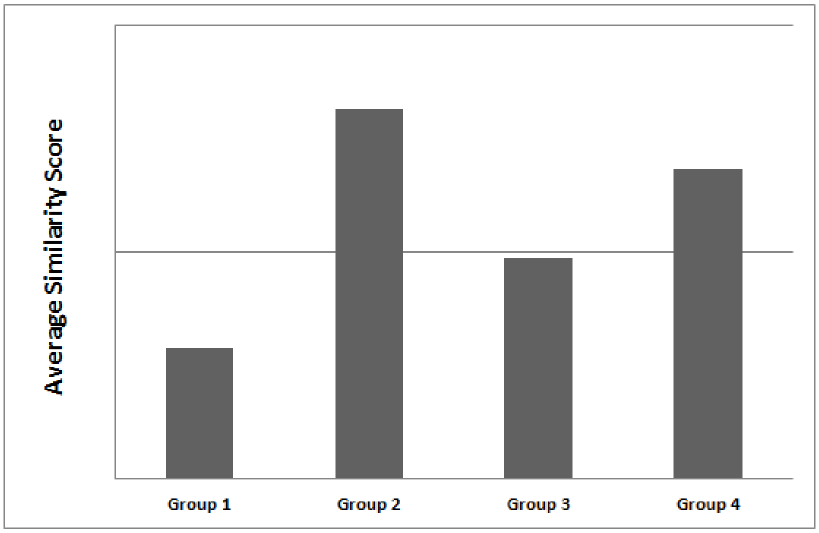
Match Rate = (Count of true positive matches + Count of true negative matches) / Count of input records

Other Matching Quality Statistics

Since the primary result of any matching run is the assignment of the input (and reference) records to matching groups, one of the basic matching quality statistics is the **distribution of record counts by matching group IDs**:

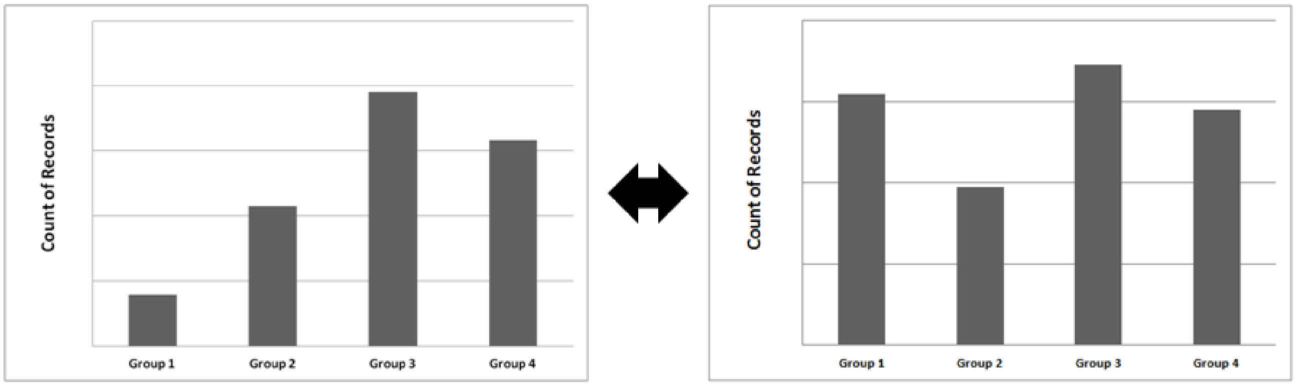


Another interesting insight in the quality of the data records assigned to the groups by the matching run could be provided by the **average similarity score associated with the records of a particular matching group**:

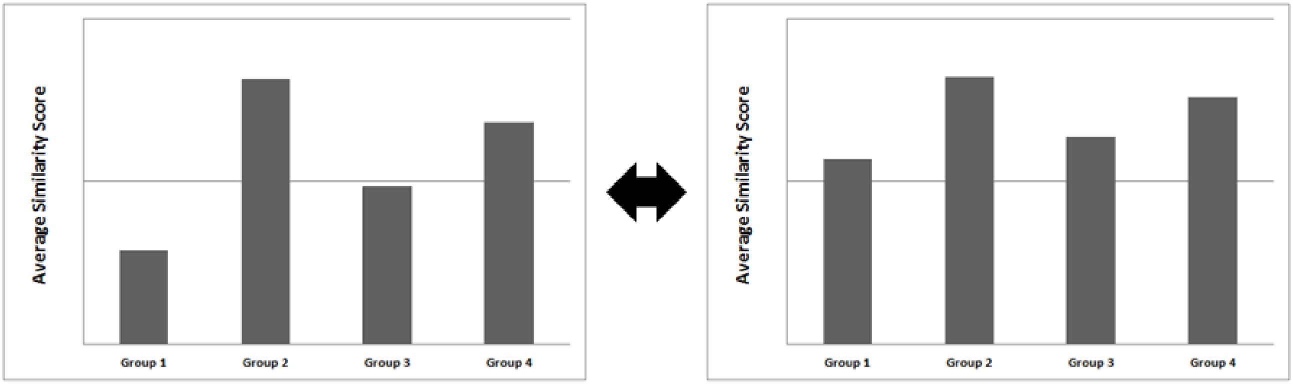


Comparison of Results between Two Matching Runs

Applying the statistics explained above, we can judge on whether or not the overall matching quality changes between the two matching runs. In the example in the below diagram, the results of the two matching runs are presented using the “count of records per matching group” statistic. We can see that the distribution of the matched records count has changed between the left and the right graphs. This change alone can hardly provide us with a clear insight in whether the matching quality improves or deteriorates, but it does signal us that the functioning of the matching scenario we are using has changed significantly (e.g., due to a change in its thresholds configuration):



A certain insight into whether the quality of our matching scenario has improved or deteriorated between the two matching runs could be obtained by comparing the graphs presenting the “average similarity score by matching” group statistic. In the example below, the graph on the left shows that the results of the first matching run have a visibly uneven distribution of the average similarity score across the matching groups. In particular, the average similarity score of Group 1 falls significantly below the average similarity score of the other three groups. If we look at the graph on the right, we will see that the average similarity score is distributed more evenly there across the matching groups. We cannot make final conclusions because we are lacking the absolute values of the average similarity scores in both graphs, but other things being equal, the graph on the right shows a more consistent matching quality than its counterpart on the left.



Related Content

SAP PIO

For more information, visit the Data Services and Data Quality Space.

Copyright

These materials are subject to change without notice. These materials are provided by SAP AG and its affiliated companies ("SAP Group") for informational purposes only, without representation or warranty of any kind, and SAP Group shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP Group products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.