



WIKIPEDIA
The Free Encyclopedia

[Main page](#)

[Contents](#)

[Featured content](#)

[Current events](#)

[Random article](#)

[Donate to Wikipedia](#)

[Wikipedia store](#)

Interaction

[Help](#)

[About Wikipedia](#)

[Community portal](#)

[Recent changes](#)

[Contact page](#)

Tools

[What links here](#)

[Related changes](#)

[Upload file](#)

[Special pages](#)

[Permanent link](#)

[Page information](#)

[Wikidata item](#)

[Cite this page](#)

Print/export

[Create a book](#)

[Download as PDF](#)

[Printable version](#)

Languages

[Français](#)

[Português](#)

[ไทย](#)

[Edit links](#)

[Create account](#) [Log in](#)

Article

[Talk](#)

[Read](#)

[Edit](#)

[View history](#)



Jaro–Winkler distance

From Wikipedia, the free encyclopedia

This article is about the measure. For other uses, see [Jaro](#).

In [computer science](#) and [statistics](#), the **Jaro–Winkler distance** (Winkler, 1990) is a measure of similarity between two [strings](#). It is a variant of the **Jaro distance** metric (Jaro, 1989, 1995), a type of string [edit distance](#), and was developed in the area of [record linkage](#) (duplicate detection) (Winkler, 1990). The higher the Jaro–Winkler distance for two strings is, the more similar the strings are. The Jaro–Winkler distance metric is designed and best suited for short strings such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match.

Contents [\[hide\]](#)

[1 Definition](#)

[2 Example](#)

[3 See also](#)

[4 References](#)

[5 External links](#)

Definition [\[edit\]](#)

The Jaro distance d_j of two given strings s_1 and s_2 is

$$d_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases}$$

Where:

- m is the number of *matching characters* (see below);
- t is half the number of *transpositions* (see below).

Two characters from s_1 and s_2 respectively, are considered *matching* only if they are the same and not farther

than $\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1$.

Each character of s_1 is compared with all its matching characters in s_2 . The number of matching (but different sequence order) characters divided by 2 defines the number of *transpositions*. For example, in comparing CRATE with TRACE, only 'R' 'A' 'E' are the matching characters, i.e. $m=3$. Although 'C', 'T' appear in both strings, they are farther than 1, i.e., $\text{floor}(5/2)-1=1$. Therefore, $t=0$. In DwAyNE versus DuANE the matching letters are already in the same order D-A-N-E, so no transpositions are needed.

Jaro–Winkler distance uses a [prefix](#) scale p which gives more favourable ratings to strings that match from the beginning for a set prefix length ℓ . Given two strings s_1 and s_2 , their Jaro–Winkler distance d_w is:

$$d_w = d_j + (\ell p (1 - d_j))$$

where:

- d_j is the Jaro distance for strings s_1 and s_2
- ℓ is the length of common prefix at the start of the string up to a maximum of 4 characters
- p is a constant **scaling factor** for how much the score is adjusted upwards for having common prefixes. p should not exceed 0.25, otherwise the distance can become larger than 1. The standard value for this constant in Winkler's work is $p = 0.1$

Although often referred to as a *distance metric*, the Jaro–Winkler distance is actually not a [metric](#) in the mathematical sense of that term because it does not obey the [triangle inequality](#) [\[1\]](#) [↗](#). In fact the Jaro–Winkler distance also does not satisfy that axiom that states that $d(x, y) = 0 \rightarrow x = y$.

In some implementations of Jaro–Winkler, the prefix bonus $\ell p (1 - d_j)$ is only added when the compared strings have a Jaro distance above a set "boost threshold" b_t . The boost threshold in Winkler's implementation was 0.7.

$$d_w = \begin{cases} d_j & \text{if } d_j < b_t \\ d_j + (\ell p(1 - d_j)) & \text{otherwise} \end{cases}$$

Example [\[edit\]](#)

Note that Winkler's "reference" C code differs in at least two ways from published accounts of the Jaro–Winkler metric. First is his use of a typo table (*adjwt*) and also some optional additional tolerance for long strings.

Given the strings s_1 MARTHA and s_2 MARHTA we find:

- $m = 6$
- $|s_1| = 6$
- $|s_2| = 6$
- There are mismatched characters T/H and H/T leading to $t = \frac{2}{2} = 1$

We find a Jaro score of:

$$d_j = \frac{1}{3} \left(\frac{6}{6} + \frac{6}{6} + \frac{6-1}{6} \right) = 0.944$$

To find the Jaro–Winkler score using the standard weight $p = 0.1$, we continue to find:

- $\ell = 3$

Thus:

$$d_w = 0.944 + (3 * 0.1(1 - 0.944)) = 0.961$$

Given the strings s_1 DWAYNE and s_2 DUANE we find:

- $m = 4$
- $|s_1| = 6$
- $|s_2| = 5$
- $t = 0$

We find a Jaro score of:

$$d_j = \frac{1}{3} \left(\frac{4}{6} + \frac{4}{5} + \frac{4-0}{4} \right) = 0.822$$

To find the Jaro–Winkler score using the standard weight $p = 0.1$, we continue to find:

- $\ell = 1$

Thus:

$$d_w = 0.822 + (1 * 0.1(1 - 0.822)) = 0.84$$

Given the strings s_1 DIXON and s_2 DICKSONX we find:[\[further explanation needed\]](#)

	D	I	X	O	N
D	1	0	0	0	0
I	0	1	0	0	0
C	0	0	0	0	0
K	0	0	0	0	0
S	0	0	0	0	0
O	0	0	0	1	0
N	0	0	0	0	1
X	0	0	0	0	0

- $m = 4$ Note that the two Xs are not considered matches because they are outside the match window of 3.
- $|s_1| = 5$
- $|s_2| = 8$
- $t = 0$

We find a Jaro score of:

$$d_j = \frac{1}{3} \left(\frac{4}{5} + \frac{4}{8} + \frac{4-0}{4} \right) = 0.767$$

To find the Jaro–Winkler score using the standard weight $p = 0.1$, we continue to find:

- $\ell = 2$







Thus:

$$d_w = 0.767 + (2 * 0.1(1 - 0.767)) = 0.814$$

See also [edit]

- [Levenshtein distance](#)
- [Record linkage](#)
- [Census](#)

References [edit]

- Cohen, W. W.; Ravikumar, P.; Fienberg, S. E. (2003). "A comparison of string distance metrics for name-matching tasks"  (PDF). *KDD Workshop on Data Cleaning and Object Consolidation* **3**: 73–8.
- Jaro, M. A. (1989). "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida". *Journal of the American Statistical Association* **84** (406): 414–20. doi:10.1080/01621459.1989.10478785 .
- Jaro, M. A. (1995). "Probabilistic linkage of large public health data file". *Statistics in Medicine* **14** (5–7): 491–8. doi:10.1002/sim.4780140510 . PMID 7792443 .
- Winkler, W. E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage"  (PDF). *Proceedings of the Section on Survey Research Methods* (American Statistical Association): 354–359.
- Winkler, W. E. (2006). "Overview of Record Linkage and Current Research Directions"  (PDF). *Research Report Series, RRS*.

External links [edit]

- [strcmp.c](#) - Original C Implementation by the author of the algorithm 

Categories: [String similarity measures](#)

This page was last modified on 17 July 2015, at 17:49.

Text is available under the [Creative Commons Attribution-ShareAlike License](#); additional terms may apply. By using this site, you agree to the [Terms of Use](#) and [Privacy Policy](#). Wikipedia® is a registered trademark of the [Wikimedia Foundation, Inc.](#), a non-profit organization.

[Privacy policy](#) [About Wikipedia](#) [Disclaimers](#) [Contact Wikipedia](#) [Developers](#) [Mobile view](#)

