

# Approximate string matching

From Wikipedia, the free encyclopedia

In computing, **approximate string matching** is the technique of finding approximate matches to a pattern in a string.

The closeness of a match is measured in terms of the number of primitive operations necessary to convert the string into an exact match. The usual primitive operations are:

- *insertion* (e.g., changing *cot* to *coat*),
- *deletion* (e.g. changing *coat* to *cot*), and
- *substitution* (e.g. changing *coat* to *cost*).

Some approximate matchers also treat *transposition*, in which the positions of two letters in the string are swapped, to be a primitive operation. Changing *cost* to *cots* is an example of a transposition.

Different approximate matchers impose different constraints. Some matchers use a single global unweighted cost, that is, the total number of primitive operations necessary to convert the match to the pattern. For example, if the pattern is *coil*, *foil* differs by one substitution, *coils* by one insertion, *oil* by one deletion, and *foal* by two substitutions. If all operations count as a single unit of cost and the limit is set to one, *foil*, *coils*, and *oil* will count as matches while *foal* will not.

Other matchers specify the number of operations of each type separately, while still others set a total cost but allow different weights to be assigned to different operations. Some matchers allow separate assignments of limits and weights to individual groups in the pattern.

Most approximate matchers used for text processing are regular expression matchers. The distance between a candidate and the pattern is therefore computed as the minimum distance between the candidate and a fixed string matching the regular expression. Thus, if the pattern is *co.l*, using the POSIX notation in which a dot matches any single character, both *coal* and *coil* are exact matches, while *soil* differs by one substitution.

The most common application of approximate matchers until recently has been spell checking. With the availability of large amounts of DNA data, matching of nucleotide sequences has become an important application. Approximate matching is also used to identify pieces of music from small snatches and in spam filtering.

## References

- *Pattern Matching Algorithms*, Alberto Apostolico & Zvi Galil, Oxford University Press, UK, 1997.

## See also

- Fuzzy string searching
- Levenshtein distance
- Needleman-Wunsch algorithm
- Soundex
- Agrep
- Zsh

# saved on 3/13/2007

Retrieved from "[http://en.wikipedia.org/wiki/Approximate\\_string\\_matching](http://en.wikipedia.org/wiki/Approximate_string_matching)"

Categories: [Articles with unsourced statements since February 2007](#) | [All articles with unsourced statements](#) | [Pattern matching](#) | [Computer science stubs](#)

- 
- This page was last modified 16:38, 14 February 2007.
  - All text is available under the terms of the GNU Free Documentation License. (See **Copyrights** for details.)
- Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a US-registered 501(c)(3) tax-deductible nonprofit charity.