



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

[Deutsch](#)
[Español](#)
[Français](#)
[Italiano](#)
[日本語](#)
[Polski](#)
[Русский](#)

 [Edit links](#)

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

Article [Talk](#)

[Read](#) [Edit](#) [View history](#)

Prediction by partial matching

From Wikipedia, the free encyclopedia
(Redirected from [PPMcompression algorithm](#))

For professional Super Smash Bros. Melee player known as PPMD, see [Kevin Nanney](#).

Prediction by partial matching (PPM) is an adaptive [statistical data compression](#) technique based on [context modeling](#) and [prediction](#). PPM models use a set of previous symbols in the uncompressed symbol stream to predict the next symbol in the stream. PPM algorithms can also be used to cluster data into predicted groupings in [cluster analysis](#).

Contents [\[hide\]](#)

- [1 Theory](#)
- [2 Implementation](#)
- [3 References](#)
- [4 See also](#)
- [5 External links](#)

Theory [\[edit\]](#)

Predictions are usually reduced to [symbol rankings](#). The number of previous symbols, *n*, determines the order of the PPM model which is denoted as PPM(*n*). Unbounded variants where the context has no length limitations also exist and are denoted as PPM*. If no prediction can be made based on all *n* context symbols a prediction is attempted with *n* − 1 symbols. This process is repeated until a match is found or no more symbols remain in context. At that point a fixed prediction is made.

Much of the work in optimizing a PPM model is handling inputs that have not already occurred in the input stream. The obvious way to handle them is to create a "never-seen" symbol which triggers the [escape sequence](#). But what probability should be assigned to a symbol that has never been seen? This is called the [zero-frequency problem](#). One variant uses the Laplace estimator, which assigns the "never-seen" symbol a fixed [pseudocount](#) of one. A variant called PPMD increments the pseudocount of the "never-seen" symbol every time the "never-seen" symbol is used. (In other words, PPMD estimates the probability of a new symbol as the ratio of the number of unique symbols to the total number of symbols observed).

Implementation [\[edit\]](#)

PPM compression implementations vary greatly in other details. The actual symbol selection is usually recorded using [arithmetic coding](#), though it is also possible to use [Huffman encoding](#) or even some type of [dictionary coding](#) technique. The underlying model used in most PPM algorithms can also be extended to predict multiple symbols. It is also possible to use non-Markov modeling to either replace or supplement Markov modeling. The symbol size is usually static, typically a single byte, which makes generic handling of any file format easy.

Published research on this family of algorithms can be found as far back as the mid-1980s. Software implementations were not popular until the early 1990s because PPM algorithms require a significant amount of [RAM](#). Recent PPM implementations are among the best-performing [lossless compression](#) programs for [natural language](#) text.

Trying to improve PPM algorithms led to the [PAQ](#) series of data compression algorithms.

A PPM algorithm, rather than being used for compression, is used to increase the efficiency of user input in the alternate input method program [Dasher](#).

References [\[edit\]](#)

- Cleary, J.; Witten, I. (April 1984). "Data Compression Using Adaptive Coding and Partial String Matching". *IEEE Trans. Commun.* **32** (4): 396–402. doi:10.1109/TCOM.1984.1096090 [↗](#).
- Moffat, A. (November 1990). "Implementing the PPM data compression scheme". *IEEE Trans. Commun.* **38** (11): 1917–1921. doi:10.1109/26.61469 [↗](#).
- Cleary, J. G.; Teahan, W. J.; Witten, I. H. (1995). "Unbounded length contexts for PPM". In Storer, J. A.;

Cohn, M. *Proceedings DCC '95*. Data Compression Conference: 28-30 Mar 1995, Snowbird, UT. [IEEE Computer Society Press](#). pp. 52–61. doi:10.1109/DCC.1995.515495 [↗](#). ISBN 0-8186-7012-6.



- C. Bloom, [Solving the problems of context modeling](#) [↗](#).
- W.J. Teahan, [Probability estimation for PPM](#) [↗](#).
- SchüRmann, T.; Grassberger, P. (September 1996). "Entropy estimation of symbol sequences". *Chaos* **6** (3): 414–427. doi:10.1063/1.166191 [↗](#). PMID 12780271 [↗](#).

See also [\[edit\]](#)

- Language model
- N-gram

External links [\[edit\]](#)

- Suite of PPM compressors with benchmarks [↗](#)
- BICOM, a bijective PPM compressor [↗](#)
- "Arithmetic Coding + Statistical Modeling = Data Compression", Part 2 [↗](#)
- (Russian)** PPMd compressor [↗](#) by Dmitri Shkarin
- PPM algorithm implementation (source code) [↗](#) by René Puchinger

v · t · e		Data compression methods	[hide]
Lossless	Entropy type	Unary · Arithmetic · Golomb · Huffman (Adaptive · Canonical · Modified) · Range · Shannon · Shannon–Fano · Shannon–Fano–Elias · Tunstall · Universal (Exp-Golomb · Fibonacci · Gamma · Levenshtein)	
	Dictionary type	Byte pair encoding · DEFLATE · Lempel–Ziv (LZ77 / LZ78 (LZ1 / LZ2) · LZJB · LZMA · LZO · LZRW · LZS · LZSS · LZW · LZWL · LZX · LZ4 · Statistical)	
	Other types	BWT · CTW · Delta · DMC · MTF · PAQ · PPM · RLE	
Audio	Concepts	Bit rate (average (ABR) · constant (CBR) · variable (VBR)) · Companding · Convolution · Dynamic range · Latency · Nyquist–Shannon theorem · Sampling · Sound quality · Speech coding · Sub-band coding	
	Codec parts	A-law · μ-law · ACELP · ADPCM · CELP · DPCM · Fourier transform · LPC (LAR · LSP) · MDCT · Psychoacoustic model · WLPc	
Image	Concepts	Chroma subsampling · Coding tree unit · Color space · Compression artifact · Image resolution · Macroblock · Pixel · PSNR · Quantization · Standard test image	
	Methods	Chain code · DCT · EZW · Fractal · KLT · LP · RLE · SPIHT · Wavelet	
Video	Concepts	Bit rate (average (ABR) · constant (CBR) · variable (VBR)) · Display resolution · Frame · Frame rate · Frame types · Interlace · Video characteristics · Video quality	
	Codec parts	Lapped transform · DCT · Deblocking filter · Motion compensation	
Theory	Entropy · Kolmogorov complexity · Lossy · Quantization · Rate–distortion · Redundancy · Timeline of information theory		
<div><div> Compression formats</div><div> Compression software (codecs)</div></div>			

Categories: Lossless compression algorithms