

Text compression

We can do this the easy way or the hard way.

2023

K. Orav

fxcm

1. Why?	3
2. Filters	4
Byte swap filter	4
3. Transforms	5
Dictionary	5
4. As fast as possible	6
5. Tuning	7
Order x	7
Byte groups for 2 and 3 bit tables	7
State tables	8
ContextMap	8
6. Contexts	9
Predictors	9
Mixers	9
mixer 0	9
mixer 1	9
mixer 2	9
mixer 3	10
mixer 4	10
mixer 5	10
mixer 6	10
mixer 7	11
mixer 8 - final mixer	11
Bracket context	11
Quote context	12
Word contexts	12
Column contexts	13
Indirect context	13
Mixed contexts	13
APM/SSE	13
Match model	14
7. Improvements	15
8. Results	16
Plain enwik8 and enwik9 compression	18
9. Future improvements	19

1. Whay?

Why not.

Result of this is the fxcn compressor. Compressor size is about 70KB (Windows executable) which can be reduced by UPX executable compressor.

Final goal was to replace paq8hp model in cmix-hp¹, that seems rather pointless at this time.

¹ <https://github.com/byronknoll/cmix-hp>

2. Filters

Filters allow the compressor to be simpler. We filter incoming data three ways:

1. Expanding size
2. Keeping size
3. Reducing size

For filtering i am using WIT². This is based on 2017 Hutter Prize winning entry filters by Alexander³. There are some changes to these filters.

I added:

- UTF8 to WC (USC) and reverse (&#xxx; to &@UTF8)
- More HTML entity swaps
- Group articles by ‘,’
- Group articles by ‘infobox’
- Group articles by number in article title
- Group articles that describe files (images, etc)

Byte swap filter

This filter is present in paq8hp12. It swaps some chars below a-z range. This is important as all chars above are related to text.

Example:

‘<’ to ‘&<’ to ‘&L’

This is important for bracket context.

In plain text mode reverse happens as byte group tables were tuned byte swap filter was on.

² <https://github.com/kaitz/wit2>

³

<https://encode.su/threads/2858-Hutter-Prize-4-17-improvement-is-here?p=58855&viewfull=1#post58855>

3. Transforms

Transforms help reduce input size. For my setup modified paq8hp12any TextFilter was used which is essentially what DRT does.

List of some programs that do text transform:

<https://github.com/kaitz/paq8pxd> (based on "XWRT 3.2 (29.10.2007)

http://www.mattmahoney.net/dc/paq8hp12any_src.zip TextFilter 3.0 for PAQ (based on WRT 4.6)

<http://nishi.dreamhosters.com/u/liptify.rar> (DRT ancestor)

<http://nishi.dreamhosters.com/u/lpaq9m.rar> (DRT)

<https://github.com/inikep/XWRT>

<http://nishi.dreamhosters.com/u/filter.rar> (Edgar Binder's preprocessor from DC)

<http://freearc.dreamhosters.com/dict.zip> (Bulat's dict preprocessor) ⁴

Dictionary

For the dictionary the version from phda 2017 November release was used.

Combined with the WIT filter this dictionary needs some changes, as some words are not present on input anymore.

⁴ <https://encode.su/threads/3240-Text-ish-preprocessors?p=62269&viewfull=1#post62269>

4. As fast as possible

I started working on paq8pxv in 2019.⁵ Biggest problem was how to implement this and what to really want from this program. There is a similar file archiver named zpaq.⁶ In zpaq you describe the model and compressor based on that input compresses data. Same is for paq8pxv.

It's slow. For paq8pxv, implementing a compression model in C(++) is about two times faster (enwik8 1500 vs 3100 sec). But there are some advantages like: no need to compile your compressor every time there is change in model, tuning parameters used for components.

So the best solution is to build a model for paq8pxv in incremental steps, tuning parameters at the same time. When it is exhausted then transfer the model to pure C code.

If needed, do manual tuning later.

⁵ <https://encode.su/threads/3064-paq8pxv-virtual-machine>; <https://github.com/kaitz/paq8pxv>

⁶ <https://mattmahoney.net/dc/zpaq.html>

5. Tuning

Most tuning was done on filtered and transformed enwik8.

Order x

First part was finding out the best order X context. I found that the order 3-5, 6,13 were the best.

In main function we input current values to BYT component:

```
for (i=0;i<7;i++) {  
    l=ci[i]; // current order  
    vmi(BYT,i,l,13,0); // set BYT component i value l with max value 13.  
}
```

In update we get our context index:

```
h=vmx(BYT, 0,0); vmx(CM, 0,t[h]);  
h=vmx(BYT, 1,0); vmx(CM, 0,t[h]);  
....
```

In tuning mode value from BYT component is in h, in every run over data it changes. If compression improves, value is kept and is output when compression finishes. So we feed this back into the model and start all over again. And again. ...

Usually about 25MB was used (megabites) for input when tuning.

Byte groups for 2 and 3 bit tables

BYT component was also used for tuning two 256 byte tables. One is named DRT_to3bits in the phda compressor, another probably DRT_to2bits.⁷ These tables are present also in paq8hp12 as WRT_mpw and WRT_mtt but only smaller. First used in paq8hp5⁸, first intended to replace direct byte parameters in the mixer.

In main:

```
// bit2word  
for (i=0;i<16;i++) {  
    x=wrt_w[i];  
    vmi(BYT,i,x,3,0);  
}  
// bit2word  
for (i=32;i<97;i++) {  
    x=wrt_w[i];  
    vmi(BYT,i,x,3,0);  
}
```

We are only interested in bytes, the first 16 and starting from space (32) until letter a(97).

⁷ <https://encode.su/threads/2769-Hutter-Prize-submission?p=52949&viewfull=1#post52949>

⁸ <http://www.mattmahoney.net/dc/text.html#1323>

In update i used 2 bit table as context:

```
w4=w4*4+wrt_w[c1];
```

```
vmx(CM, 7,w4);
```

Tuning was done on full input. Same was done on the wrt_t table (bit3word).

State tables

At the start there were improvements when tuning. Later really small compression improvements. ContextMap uses RND function to random transition state to another when the state reaches 204 (this is used by all paq8 versions). I tested multiple different ways how state tables were generated. I also wanted to remove random transitions.

Removed part where *States 0-30 are a history of the last 0-4 bits* and when generating table i disabled *state<15*. Discount was also changed to a tunable parameter. Finally there were compression improvements and also speed improvements when random state transitions are removed.

All state tables that improved compression but contained unreachable states were discarded. Sometimes manual parameter change was needed to reach 255 states.

Final result was four different state tables. There were multiple reruns as contexts were added.

I think there is still room for improvement.

ContextMap

There are four main parameters in ContextMap. Each one is for one prediction. These values are used to calculate prediction tables so when compressing we map prediction or current state to these tables.

Most values were tuned.

6. Contexts

Predictors

There are a total of 62 contexts to ContextMap, 1 context to RunContextMap, 8 SmallStationaryContextMap contexts, 3 StateMap contexts 1 direct prediction for match model. This adds up to 396 predictions to mixers.

Contexts were incrementally added then tested and tuned. At the same time state tables were tuned.

Only APM contexts are hashed. Hashing hurts compression most of the time is used for ContextMap contexts. Also adds to code size and degrades compression time. I wanted to keep the compression speed between 32-40 kb /s on my test computer.

I think 5-10 contexts for ContextMap can be removed or replaced with better ones. It is hard task.

Mixers

There are 9 mixers. First 8 produce predictions to the final mixer (9). Each mixer has context. Here are the contexts:

mixer 0

- context is sum of context order(3-5,6,8) isState counts (max 6) and bit pos

mixer 1

- at bpos=0 context is last 2 bit2word and 1 bit3word
- at bpos=1-3 context is last 2 bit2word and 1 bit3word of current bracket or quote
- at bpos=4-7 context is last 1 bit2word, current bit2word from c0 and bit3word of current bracket or quote

mixer 2

- at bpos=0 context is was byte(3,4) a word and 2 bit3word
- at bpos=1 context is bit 1xxxxxxx from c0, was byte(2) a word, bit pos max 5,last 1 bit3word and 1 bit3word of current bracket or quote
- at bpos=2 context is bit 11xxxxxx (bit pos 2) from c0, was byte(2) a word, bit pos max 5,last 1 bit3word and 1 bit3word of current bracket or quote
- at bpos=3 context is bit 111xxxxx (bit pos 3) from c0, was byte(2) a word, bit pos max 5,last 1 bit3word and 1 bit3word of current bracket or quote
- at bpos=4-7 context is bit current bit2word from c0, bit pos max 5,last 1 bit3word and 1 bit3word of current bracket or quote

mixer 3

- at bpos=0-7 context is was byte(3,4) a word, sum of context order(3-5,6,8) isState counts (max 6) and last 2 bit2word

mixer 4

- at bpos=0 context is bit xxxxxxxx from c0, was byte(1-8) a word or space and bit pos
- at bpos=1 context is bit 1xxxxxxx from c0, was byte(1-7) a word or space and bit pos
- at bpos=2 context is bit 11xxxxxx from c0, was byte(1-6) a word or space and bit pos
- at bpos=3 context is bit 111xxxxx from c0, was byte(1-5) a word or space and bit pos
- at bpos=4 context is bit 1111xxxx from c0, was byte(1-4) a word or space and bit pos
- at bpos=5 context is bit 11111xxx from c0, was byte(1-3) a word or space and bit pos
- at bpos=6 context is bit 111111xx from c0, was byte(1-2) a word or space and bit pos
- at bpos=7 context is bit 1111111x from c0, was byte(1) a word or space and bit pos

mixer 5

- at bpos=0 context is bit xxxxxxxx from c0, first char type state(0,1) xxxx1xxx, 2 bit2word 1111xxxx
- at bpos=1 context is bit 1xxxxxxx from c0, first char type state(0,1) xxxx1xxx, 1 bit3word x111xxxx, bit pos xxxxx111
- at bpos=2 context is bit 11xxxxxx from c0, first char type state(0,1) xxxx1xxx, 1 bit2word xx11xxxx, bit pos xxxxx111
- at bpos=3 context is bit 111xxxxx from c0, first char type state(0,1) xxxx1xxx, was byte(1) a word xxx1xxxx, bit pos xxxxx111
- at bpos=4 context is bit 1111xxxx from c0, first char type state(0,1) xxxx1xxx, bit pos xxxxx111
- at bpos=5 context is bit 11111xxx from c0, first char type state(0,1) xxxx1xxx (overflow, ok!)
- at bpos=6 context is bit 111111xx from c0, first char type state(0,1) xxxx1xxx
- at bpos=7 context is bit 1111111x from c0, first char type state(0,1) xxxx1xxx
- at bpos=0-7 sum of context order(3-5,6,8) isState counts (max 5) and is match(0,1) 111 xxxxxxxx

mixer 6

- at bpos=0-7 context is sum of context words isState counts (max 6), first char type state(0,1), 2 bit2word of byte(3,4) and 1 bit3word of byte(2)

mixer 7

- at bpos=0-7 context is non-repeating 2 bit3word of byte(2,3), was byte(1-3) a word and last 1 bit2word

mixer 8 - final mixer

- at bpos=0-7 context is sum of context order(3-5,6,8) isState counts (max 6), bracket or quote state(0,1) and last 1 bit2word

There may be room for more mixer contexts.

Bracket context

In cmix there is BracketContext⁹. It is good but we can do better. First problem is it uses modern C++ witch to me seems overkill, there is code bloat. Also context is always updated. What we should do is update only when we see non letter char so our distance to the last bracket stays in range.

Same is not true for the plain text version, in that case we need to include all letters (upper/lower case) and extend bracket length limit. When this extra case is not used we lose about 100 bytes per 1 mb.

It seems this context is really important for enwik8/enwik9.

```
<minor />
<comment>Fixing redirect</comment>
<text xml:space="preserve">#REDIRECT [[accessible computing]]</text>
</revision>
</page>
</page>
<title>Anarch</title>
<id>11</id>
</revision>
<id>15898945</id>
<timestamp>2002-09-22T16:02:58Z</timestamp>
<contributor>
  <username>Andre Engels</username>
  <id>300</id>
</contributor>
</minor />
<text xml:space="preserve">#REDIRECT [[ada programming language]]</text>
</revision>
</page>
</page>
<title>Anarchism</title>
<id>12</id>
</revision>
<id>42136831</id>
<timestamp>2006-03-04T01:41:25Z</timestamp>
<contributor>
  <username>CJames745</username>
  <id>832382</id>
</contributor>
</minor />
<comment># Anarchist Communism # too many brackets</comment>
<text xml:space="preserve">
'''Anarchism''' originated as a term of abuse first used against early [[working class]] [[radicals]] including the [[Diggers]] of the [[English Revolution]] and the [[sans-culottes]] of the [[French Revolution]].[http://uk.encyclopedia.msn.com/encyclopedia/61568720/Anarchism.html] Whilst the term is still used in a pejorative way to describe ''&quot;any act that used violent means to destroy the organization of society&quot;;&#39;&#39; it is also used to describe a political philosophy. In the [[English language]] it is used to describe a political philosophy that advocates the elimination of authoritarian institutions, particularly the [[state]].[http://en.wikipedia.org/wiki/Definitions_of_anarchism] Definitions of anarchism on Wikiquote, accessed 2006&#39;&#39; The word &quot;[[anarch]]&quot; as most anarchists use it, does not imply [[chaos]], [[inhibition]], or [[anomie]], but rather a harmonious [[anti-authoritarian]] society. In place of what are regarded as authoritarian political structures and coercive economic institutions, anarchists advocate social relations based upon [[voluntary association]] of autonomous individuals, [[mutual aid]], and [[self-governance]].

While anarchism is most easily defined by what it is against, anarchists also offer positive visions of what they believe to be a truly free society. However, it is also about how an anarchist society might work vary considerably, especially with respect to economics; there is also disagreement about how a free society might be brought about.

== Origins and predecessors ==

[[Peter Kropotkin|Kropotkin]], and others, argue that before recorded [[history]], human society was organized on anarchist principles.&#39;&#39;[[Peter Kropotkin|Kropotkin]], Peter. ''&quot;[[Mutual Aid: A Factor of Evolution]]&quot;;'', 1902.&#39;&#39; Most anthropologists follow Kropotkin and Engels in believing that hunter-gatherer bands were egalitarian and lacked division of labour, accumulated wealth, or decreed law, and had equal access to resources.&#39;&#39;[[Friedrich Engels|Engels]], Friedrich. ''&quot;[http://www.marxists.org/archive/marx/works/1884/origin-family/index.htm] Origins of the Family, Private Property, and the State&quot;;'', 1884.&#39;&#39;&#39;

[[Image:WilliamGodwin.jpg|thumb|right|150px|William Godwin]]

Anarchists including the [[The Anarchy Organisation]] and [[Murray Rothbard|Rothbard]] find anarchist attitudes in [[Taoism]] from [[History of China|Ancient China]].&#39;&#39;&#39;The Anarchy Organization (Anonot). ''Taoism and Anarchy''. [[Q&A]] [[2002]] [http://www.toradepop.co.uk/~anarch/taoism.htm] toradepop mirrow [http://www.geocities.com/Solo/5705/taoan.html] Unity site mirrow&#39;&#39;&#39;[[Murray Rothbard|Rothbard]], Murray. ''&quot;[http://www.lewrockwell.com/rothbard/ancient-chinese.html] The Ancient Chinese Libertarian Tradition&quot;;'', an extract from ''&quot;[http://www.mises.org/journals/jls/2/2/2.3.pdf] Concepts of the Role of Intellectuals in Social Change Toward Leisure&quot;;'', The Journal of Libertarian Studies, 2 (2) Fall 1990.&#39;&#39;&#39;[[Peter Kropotkin|Kropotkin]] found similar ideas in [[Stoicism|Stoic]] [[Zeno of Citium]]. According to Kropotkin, Zeno &quot;repudiated the omnipotence of the state, its intervention and regimentation, and proclaimed the sovereignty of the moral law of the individual&quot;.&#39;&#39;&#39;[[http://www.blackcrayon.com/page.jsp?libreary/ncit1910.html] Anarchism], written by Peter Kropotkin, from Encyclopaedia Britannica, 1910&#39;&#39;&#39;

The [[Anabaptists]] of 16th century Europe are sometimes considered to be religious forerunners of modern anarchism. [[Bertrand Russell]], in his ''History of Western Philosophy'', writes that the Anabaptists &quot;repudiated all law, since they held that the good man will be guided at every moment by [[the Holy Spirit]]... If from this premise they arrive at [[communism]]...&quot;&#39;&#39;[[Bertrand Russell|Russell]], Bertrand. ''&quot;Ancient philosophy&quot;;'', in ''A History of Western Philosophy, and its connection with political and social circumstances from the earliest times to the present day'', 1945.&#39;&#39;&#39;[[Diggers|True Levellers|The Diggers]] or &quot;True Levellers&quot;; were an early communist movement during the time of the [[English Civil War]], and are considered by some as forerunners of modern anarchism.&#39;&#39;&#39;[[http://www.gpub.com/notes/aan-hist.html] An Anarchist Timeline], from Encyclopaedia Britannica, 1994.&#39;&#39;&#39;

In the [[modern era]], the first to use the term to mean something other than chaos was [[Louis-Armand de Laon d'Arce de Lahontan, Baron de Lahontan|Louis-Armand de Lahontan]] in his ''Nouveaux voyages dans l'Amérique septentrionale'', (1703), where he described the [[Native Americans in the United States|indigenous American society]], which had no state, laws, prisons, priests, or private property, as being in anarchy&#39;&#39;&#39;[[http://text.lib.queensia.edu/cgi-local/DNI/dhi.cgi?id=dvi-12] Dictionary of the History of Ideas - ANARCHISM&#39;&#39;&#39;The word &quot;anarchy&quot; was first used in the [[American Indian Movement]] in 1973, in the thick of the [[French Revolution]]. [[William Godwin]] published ''An Enquiry Concerning Political Justice'' [http://web.bilkent.edu.tr/Online/
```

⁹ <https://github.com/byronkholl/cmix/blob/master/src/contexts/bracket-context.cpp>

Quote context

Bracket context can be used also for quotes. There are some differences. We need to escape some char combinations or compression will be worse.

Quote is removed when following conditions are met:

- "x'x" where x is any letter in word
- "]'x" where x is any letter in word
- "y'x" where y is number and x is any letter in word
- "x' " where x is any letter in word
- "x'@" where x is number
- "x'@" where x is any letter in word

```
<comment>Automated conversion</comment>
<text xml:space="preserve">#REDIRECT [[Applied ethics]]</text>
</revision>
</page>
<page>
<title>Accessible Computing</title>
<id>10</id>
<revision>
<id>15878945</id>
<timestamp>2003-04-25T22:18:38Z</timestamp>
<contributor>
<username>Ams80</username>
<id>7543</id>
</contributor>
<minor />
<comment>Fixing redirect</comment>
<text xml:space="preserve">#REDIRECT [[accessible_computing]]</text>
</revision>
</page>
<page>
<title>Ada</title>
<id>11</id>
<revision>
<id>15878946</id>
<timestamp>2002-09-22T16:02:58Z</timestamp>
<contributor>
<username>Andre Engels</username>
<id>300</id>
</contributor>
<minor />
<text xml:space="preserve">#REDIRECT [[ada programming language]]</text>
</revision>
</page>
<page>
<title>Anarchism</title>
<id>12</id>
<revision>
<id>42136831</id>
<timestamp>2006-03-04T01:41:25Z</timestamp>
<contributor>
<username>James745</username>
<id>832382</id>
</contributor>
<minor />
<comment>/* Anarchist Communism */ too many brackets</comment>
<text xml:space="preserve"><(Anarchism)>
'''Anarchism'' originated as a term of abuse first used against early [[working class]] and the ''sans-culotte'' &quot;sans-culottes'' of the [[French Revolution]]. It has since been used in a pejorative way to describe '''&quot;any act that used violent means to destroy cy/faculty/deflen/zhistoirntopolency.html History of International Police Cooperation]]. The Social Defense Against Anarchists&quot;; 1898&lt;/ref&gt;; it has also been taken
```

Word contexts

Dictionary transform adds escape char to all bytes above 127. Most word contexts are set only when the previous byte is not escaped char. Some context are:

word(0)

```
word(0)+word(1)+lastByte
```

word(0)+word(2)+lastByte

```
word(0)+word(2)+word(3)+lastByte
```

word(0)+firstWord+lastByte

word(1)-word(0)+word(2)

```

<page>
<title>Anarchism</title>
<id>12</id>
<revision>
<id>42136831</id>
<timestamp>2006-03-04T01:41:25Z</timestamp>
<contributor>
<username>CJames745</username>
<id>832382</id>
</contributor>
<minor />
<comment>/* Anarchist Communism */ too many brackets</comment>
<text xmlns="preserve">X<Anarchism>
'''Anarchism''' originated as a term of abuse first used against early [[working class]] [[
d the [[sans-culottes|'sans-culottes']] of the [[French Revolution]].http://uk.encyclopedia.com
used in a pejorative way to describe '''&quot;any act that used violent means to destroy th
cy/faculty/deflem/zhistorintpolency.html History of International Police Cooperation]], from
r the Social Defense Against Anarchists&quot;, 1898&lt;/ref&gt;, it has also been taken up
The word '''anarchism''' is \[etymology|derived from\] the \[\[Greek language|Greek\]\] '''\[Wik
&#945;,&#945;,&#957;,&#945;,&#961;,&#967;,&#943;,&#945;\]\]''' <quot
litical philosophy]], is the belief that '''rulers''' are unnecessary and should be abolished
archism also refers to related \[\[social movement\]\]s that advocate the elimination of autho
en.wikiquote.org/wiki/Definitions_of_anarchism Definitions of anarchism on Wikiquote, acce
hists use it, does not imply \[\[chaos\]\], \[\[nihilism\]\], or \[\[anomie\]\], but rather a harmoniou
thoritarian political structures and coercive economic institutions, anarchists advocate so
ividuals, \[\[mutual aid\]\], and \[\[self-governance\]\].
While anarchism is most easily defined by what it is against, anarchists also offer positiv
deas about how an anarchist society might work vary considerably, especially with respect t
t be brought about.
== Origins and predecessors ==
[[Peter Kropotkin|Kropotkin]], and others, argue that before recorded \[\[history\]\], human so
kin:Kropotkin]], Peter. '''&quot;\[\[Mutual Aid: A Factor of Evolution\]\]&quot;''', 1902.&lt;/re
bat hunter-gatherer bands were egalitarian and lacked division of labour; accumulated wealth
edrich Engels|Engels]], Friedrich. '''&quot;http://www.marxists.org/archive/marx/works/1884
the State&quot;''', 1884.&lt;/ref&gt;
[[Image:WilliamGodwin.jpg|thumb|right|150px|William Godwin]]

```

Column contexts

For column contexts we keep track of column length, above byte and previous above byte.

Some context are:

above +col

above+prevabove+col+lastByte

above+prevabove+lastTowBytes

col+firstChar+bracket+bit3word(2,1)

Indirect context

There are two indirect contexts. One byte and two byte one. First one also depends on what is the current bracket context.

Mixed contexts

There are many mixed contexts. Some use byte-bit groups, columns, word, bytes and any combination of them.

APM/SSE

In the tuning phase no APMs were present. There was an unaligned bitmodel UAS that used prediction error as context. This was later used for APMs and the UAS component was removed. APMs are the same as in paq8hp12 but with some changes to contexts.

Match model

Match model is mostly identical to one in pax8px (v209). Maximum match length is limited to 62 bytes.

```
<siteinfo>
  <sitename>Wikipedia</sitename>
  <base>http://en.wikipedia.org/wiki/Main_Page</base>
  <generator>MediaWiki 1.6alpha</generator>
  <case>first-letter</case>
  <namespaces>
    <namespace key="-2">Media</namespace>
    <namespace key="1">Special</namespace>
    <namespace key="0" />
    <namespace key="1">Talk</namespace>
    <namespace key="2">User</namespace>
    <namespace key="3">User talk</namespace>
    <namespace key="4">Wikipedia</namespace>
    <namespace key="5">Wikipedia talk</namespace>
    <namespace key="6">Image</namespace>
    <namespace key="7">Image talk</namespace>
    <namespace key="8">MediaWiki</namespace>
    <namespace key="9">MediaWiki talk</namespace>
    <namespace key="10">Template</namespace>
    <namespace key="11">Template talk</namespace>
    <namespace key="12">Help</namespace>
    <namespace key="13">Help talk</namespace>
    <namespace key="14">Category</namespace>
    <namespace key="15">Category talk</namespace>
    <namespace key="100">Portal</namespace>
    <namespace key="101">Portal talk</namespace>
  </namespaces>
</siteinfo>
<page>
  <title>AaA</title>
  <id>1</id>
  <revision>
    <id>32899315</id>
    <timestamp>2005-12-27T18:46:47Z</timestamp>
    <contributor>
      <username>Jsmethers</username>
      <id>614213</id>
    </contributor>
    <text xml:space="preserve">#REDIRECT [[AaA]]</text>
  </revision>
</page>
<page>
  <title>Algeria</title>
  <id>5</id>
  <revision>
    <id>18063769</id>
    <timestamp>2005-07-03T11:13:13Z</timestamp>
    <contributor>
      <username>Docu</username>
      <id>8029</id>
    </contributor>
    <minor />
    <comment>adding cur_id=5: <<R from CamelCase>></comment>
    <text xml:space="preserve">#REDIRECT [[Algeria]]<<R from CamelCase>></text>
  </revision>
</page>
<page>
  <title>American Samoa</title>
  <id>6</id>
  <revision>
```

7. Improvements

I call these as improvements as they improved compression rate and time.

ContextMap

1. In ContextMap use prebuilt prediction tables.
2. Remove RND state update.
3. Tuned state tables.
4. In hash table when replacing lowest priority value keep last accessed value for some contexts

Mixer

1. When last byte prediction had no error compared to our thresholds, increase the limit when we update mixer weights.

Bracket context

1. Modified to provide better context, which improves compression.

8. Results

Test computer parameters: Intel Core i5 4460 3.2GHz, DDR3 PC3-12800 (800MHz) 32GB.

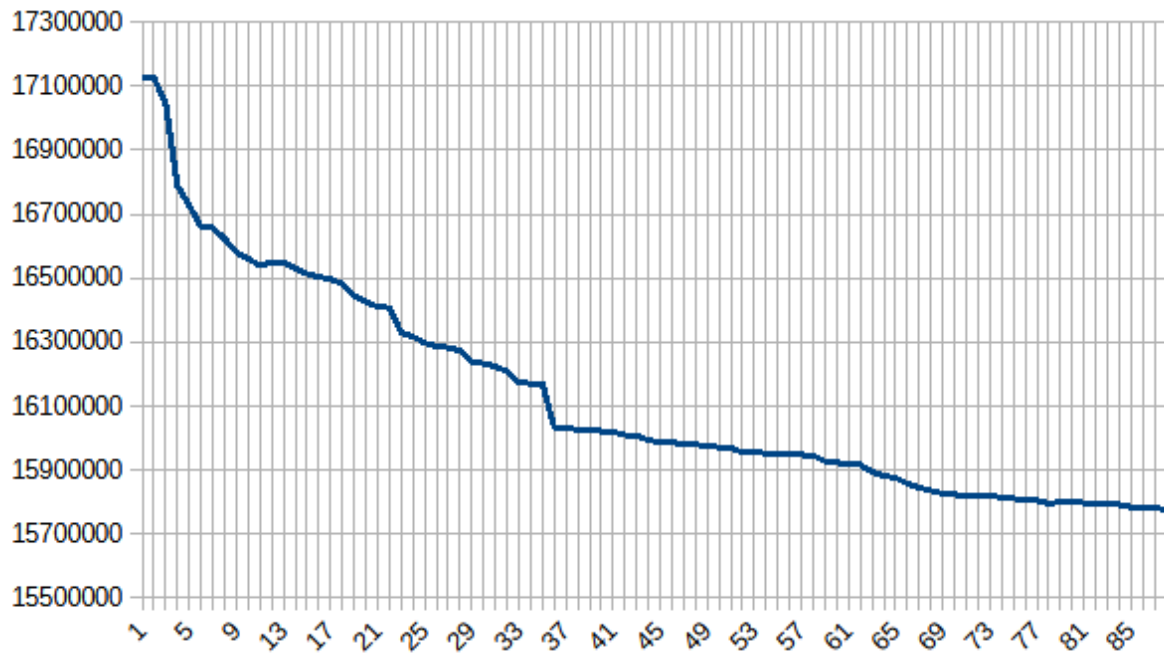
DIC - dictionary transform: paq8hp12 -0.

WIT - <https://github.com/kaitz/wit2>

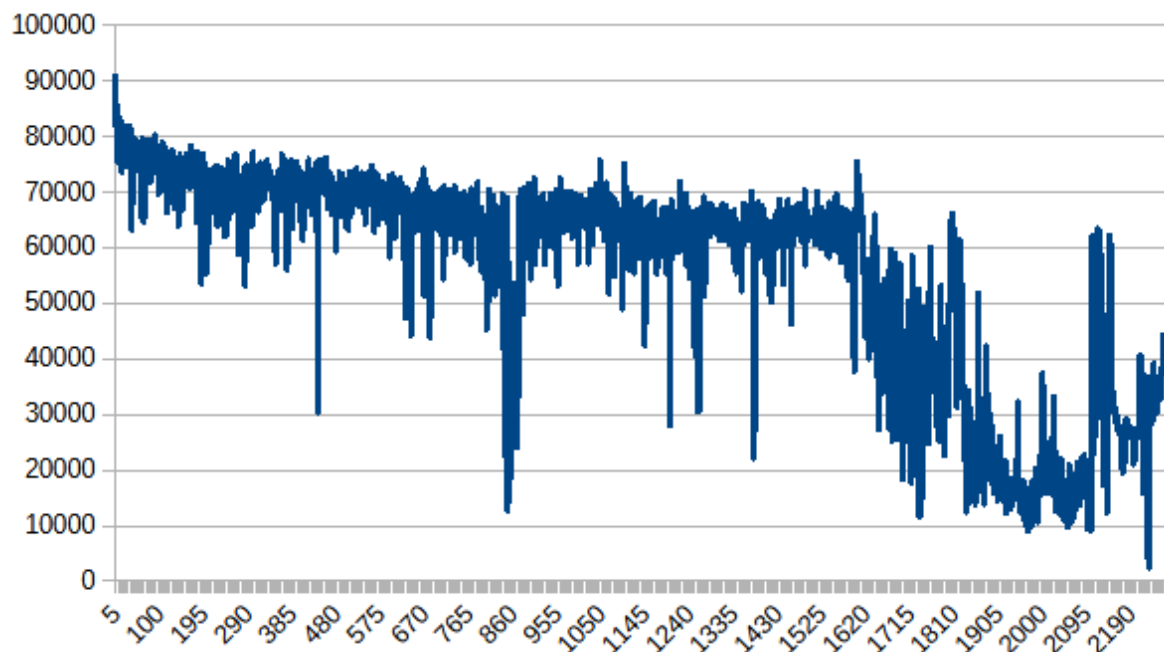
	File	Transforms	Size	Compressed	Time sec	Memory
MB						
fxcm v1 1840	enwik8	-	100000000	16675996	2934	
fxcm v1 1840	enwik9	-	1000000000	135192577	26322	
fxcm v2 1840	enwik8	WIT,DIC	57088865	15761972	1880	
fxcm v2 1840	enwik9	WIT,DIC	593869820	126234551	17121	
paq8pxd_v107 -s7 1460	enwik8	DIC (internal)	100000000	16408142	11189	
paq8pxd_v107 -s8 2264	enwik8	DIC (internal)	100000000	16182108	11473	
paq8n -8 1567	enwik8	-	100000000	17916450	5663	
paq8n -8 1567	enwik8	WIT,DIC	57088865	16905680	3457	
paq8o10t -8 1517	enwik8		100000000	17772821	6017	
paq8o10t -8 1517	enwik8	WIT,DIC	57088865	17101300	2914	
paq8px_v208fix1 -8 1813	enwik8		100000000			
paq8px_v208fix1 -8 1813	enwik8	WIT,DIC	57088865			
paq8hp12any 1813	enwik8	WIT,DIC	57088865	16131394	2393	
paq8hp12any 1813	enwik9	WIT,DIC	593869820	130573629	24396	

paq8n, paq810t are original compilations from LTCB, for speed comparison. paq8hp12any has dictionary transform removed, mixer has AVX2 train and update. paq8px_v208fix1 has an extra option -forcetext for processed enwik8. Dictionary is compressed by fxcm 460288 -> 79211.

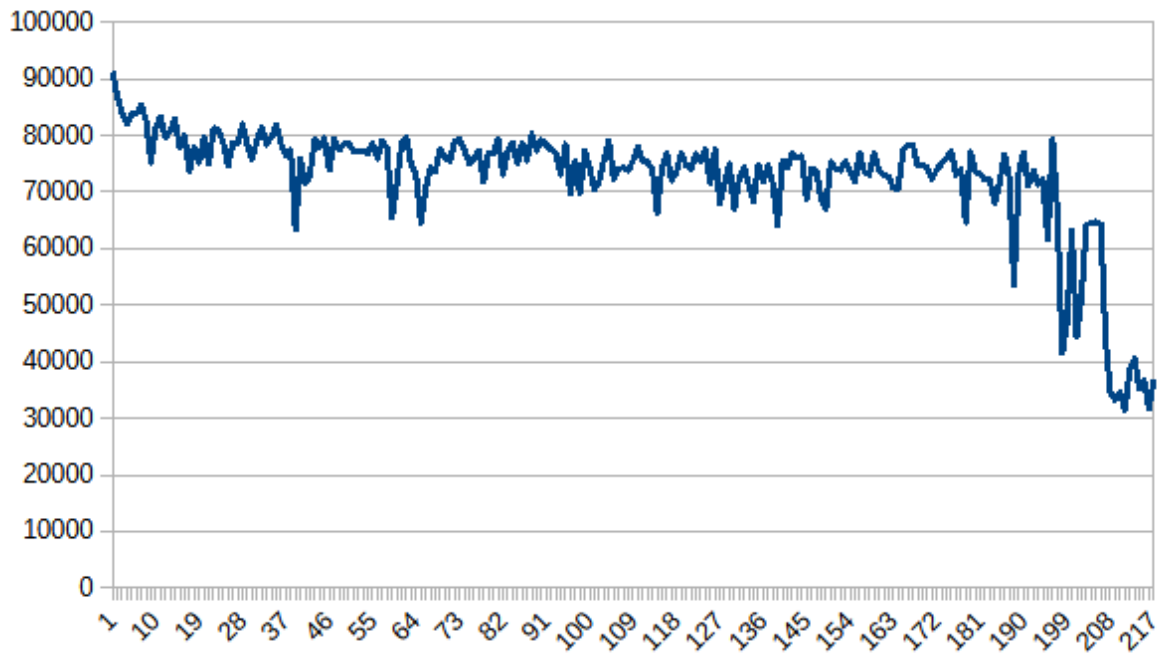
There are about 55 different stable versions of tuned versions of fxcm in paq8pxv config format. I lost track of how many tuning rounds there were. After porting the config file to C++ code there were about 56 stable test versions.



Compression improvements over time for enwik8 (WIT, DIC). Drop around version 37 is when APMs were added.



Compression of processed enwik9 (WIT, DIC). Compressed size 593869820 -> 126234551 bytes (not including compressed dictionary), time about 17121 sec.



Compression of processed enwik8 (WIT, DIC). Compressed size 57088865 -> 15761972 bytes (not including compressed dictionary), time about 1880 sec.

Memory usage is about 1840 MB. Compression speed is around 32-40 kb depending on the input.

Above results are as of 27.08.2023.

Plain enwik8 and enwik9 compression

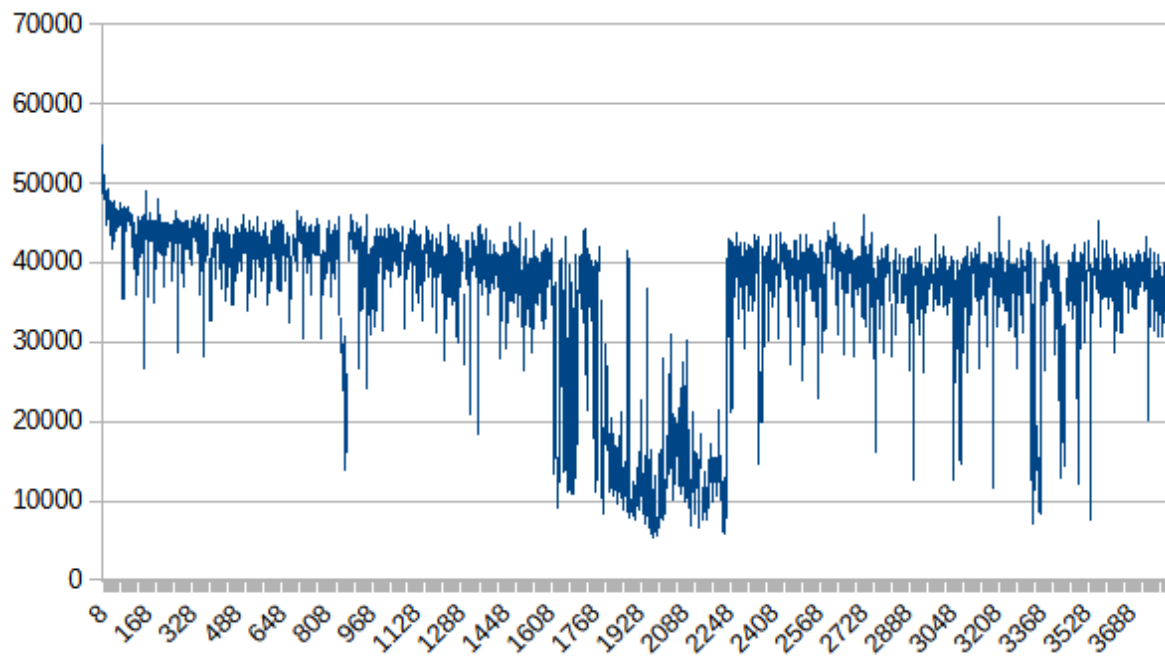
According to LTCB¹⁰ this is the best compression ratio for the non dictionary version. Just above zpaq (11).

Compressor fxcv1:

100000000 -> 16675996 in 2934 sec, 1829 MB RAM.

1000000000 -> 135192577 in 26322 sec, 1829 MB RAM

¹⁰ <http://www.mattmahoney.net/dc/text.html>



enwik9 compression ratio.

9. Future improvements

- ~~Deferred update in ContextMap as in paq8px version.~~ Tested 18.08.2023 - worse or some mistake in code.
- Table based mask for 2 and 3 bit tables based on first char.
- Better match model. - 22.08.2023 added matchmodel from paq8px, it is better.
- ~~3-layer mixer.~~ As of 22.08.2023 does not seem to work, or really small improvements.
- Better hash replacement in ContextMap, extend improvement 4.
- Modified dictionary. word2vec
- Decode dictionary codewords back to text, as in paq8pxd
- Use decoded dictionary words with wordmodel/stemmer.
- Different APM/SSE - *not likely*.