圧縮を用いた類似度判定のための計算実験

Experiments for compression based similarity distance

谷　研究室　　森田　岳史
Takeshi Morita

概要

圧縮に基づいたデータ間の類似度に関する距離が定義され,DNA、言語、音楽の類似度判定に有用だという実験結果が得られている. 本研究ではその類似度判定のための計算実験を行う。

1　はじめに

近年、急速なコンピュータの発達によりの様々な事柄（音楽や文章など）が電子的データとして扱われるようになった。Ming Li，Xin Chen，Bin Ma，Paul M.B. Vitan yi,Rudi Cilibrasi らは Kolmogorov 記述量に基づいたデータ間の距離を表す NID(Normalized Information Distance) を提案した。その NID は万能性を持ち、あらゆるデータに対して適用できデータそのものの前提や知識を必要としないことから similarity metric と呼ばれている。ただ、Kolomogorov 記述量は計算不可能であることが知られている。そこで、NID を計算可能にするため圧縮を用いた NCD(normalized compression distance）が提案された (NCD より DNA の類似度や言語の類似度、音楽の類似度判定に有用だということが分かっている。)　本研究では圧縮に基づいた類似度判定のための計算実験を行い、類似度判定に優れる圧縮アルゴリズムについて、考察する。第 2 節では Kolmogorov 記述量について，第 3 節では NID および NCD について述べ，第 4 節では実験で用いるノーマル圧縮 (Noramal Compressor) について述べ、第 5 節では、計算実験を行い、結果を考察する。第 6 節では、今後の課題を述べる。

2　Kolmogorov 記述量の定義について

この節では Kolmogorov 記述量の基本的な定義について、述べる。

2.1　Kolmogorov 記述量

定義

あるデータ x が存在し、データ x の Kolmogorov 記述量とはあるプログラム言語で x を生成する最小のプログラムのサイズである。どんなプログラム言語を選んでも、それが妥当であれば情報量は定数分の差しかないことが知られている。S をプログラム言語、$|p|$ をプログラムサイズとすると、データ x の Kolmogorov 記述量 $K_s(x)$ は、以下のように定義される。

$$K_s(x) = min\{|p| : S(p) = x\}$$

次に補助情報 y に対するデータ x の Kolomogorov 記述量 $k(x|y)$ を U はデータを記述する言語として固定した万能言語であるとして以下と定義する。

$$K(x|y) = min\{|p| : U(p, y) = x\}$$

2.2　Kolmogorov 記述量のアルゴリズム的性質

$K(x)$ を正の整数から正の整数への関数と考える。

定理

(a) 関数 $K(x)$ は帰納的ではない。しかも、どのような機能的部分関数 $\phi(x)$ を考えても、もしその値が無限個の点で定義されているのならば、その定義上のどこかの点で $\phi(x) \neq K(x)$ となる。
(b) 引数 t に対しては単調減少で（全域的な）機能的関数 $H(t, x)$ が存在し,$\lim_{t\to\infty} H(t, x) = K(x)$。すなわち、$K(x)$ の良い近似は計算可能。ただし一様近似ではない。

3　ＮＩＤおよびＮＣＤについて

3.1　距離

数学において距離空間とは，任意の 2 点間で距離が定められた空間のことをいう.

定義

ある集合 X 上の距離とは，実数値関数 $d : X \times X \to R$ で任意の $x, y, z \in X$ に対して次のような性質を満たす。

$$d(x, y) \geq 0$$
$$d(x, y) = 0 \Leftrightarrow x = y$$
$$d(x, y) = d(y, x) : \text{対称性}$$
$$d(x, y) \leq d(x, z) + d(z, y) : \text{三角不等式}$$

3.2 NID(Normalized Information Distance)

Ming Li らの研究では，情報に関する距離を標準化し、その距離をＮＩＤと呼ぶ。

定義

任意のデータ x, y について，以下のように決める。ＮＩＤを定義する。

$$NID(x,y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}}$$

3.3 NCD(Normalized Compression Distance)

kolmogorov 記述量は計算することすることが出来ない。K の近似として圧縮アルゴリズムを用いた類似度距離を NCD と呼ぶ。

定義

文字列 x をある圧縮アルゴリズム C で圧縮したサイズを C(x)、x と y の連接を xy と表すとする。

$$NCD(x,y) = \frac{C(xy) - min\{C(x), C(y)\}}{max\{C(x), C(y)\}}$$

また，$C(y) \geq C(x)$ としたとき，

$$NCD(x,y) = \frac{C(xy) - C(x)}{C(y)}$$

NCD は０以上１以下の値域をとり、０のときに類似度が高く、１のとき類似度が低いと考える。

4 ノーマル圧縮 (Normal Compressor)

ＮＣＤで使用される圧縮アルゴリズムは万能性を保つためにいくつかの制約が必要とされる。以下ではＮＣＤにおいて、望まれる圧縮アルゴリズムについての条件を定義する。

定義

x:任意のデータファイル

C:圧縮機 (Normal Compressor)

$C(x)$:ファイル x を圧縮機 C で圧縮した後のファイルサイズとして

$$C(\lambda) = 0 \Leftrightarrow \lambda\text{が空のファイル}$$
$$C(xx) = C(x)$$
$$C(xy) \geq C(x)$$
$$C(xy) = C(yx)$$
$$C(xy) \leq C(x) + C(y)$$
$$C(xy) + C(z) \leq C(xz) + C(yz)$$

以上は、理想されるもので、若干の誤差は許容される。

5 計算実験

5.1 実際の圧縮機で条件について適合度実験

gzip と bzip2 の圧縮アルゴリズムでデータファイルを圧縮し、どちらのアルゴリズムがノーマル圧縮の定義を満たしているか比較実験する。

実験

使用する圧縮アルゴリズム：bzip2,gzip

実験方法：桃太郎方言のファイルを使用し、定義に当てはめた圧縮後のサイズを記録する。

以下が結果の数値である。

5.1.1 方言桃太郎の都道府県別テキストファイル

	$C(xx)$	$=$	$C(x)$
gzip	282.178	:	300.464
bzip2	231.982	:	280.785
	$C(xy)$	$\geq$	$C(x)$
gzip	467.472	:	282.690
bzip2	370.133	:	231.982
	$C(xy)$	$\leq$	$C(x) + C(y)$
gzip	467.472	:	$282.690 + 281.666$
bzip2	370.133	:	$232.346 + 231.617$
	$C(xy)$	$=$	$C(yx)$
gzip	467.472	:	467.567
bzip2	371.758	:	371.754
	$C(xy) + C(z)$	$\leq$	$C(xz) + C(yz)$
gzip	444.7+282.266	:	$458.333 + 461.033$
bzip2	370.714+232	:	$385.607 + 385.142$

(単位：*byte*　値はすべて平均)

5.1.2 同研究室、荒堀、福田よる 4.1.1 のファイルの前処理 1

	$C(xx)$	$=$	$C(x)$
gzip	193.107	:	204.892
bzip2	205.589	:	246.107
	$C(xy)$	$\geq$	$C(x)$
gzip	320.859	:	193.607
bzip2	344.503	:	206.584
	$C(xy)$	$\leq$	$C(x)+C(y)$
gzip	320.859	:	$193.607+192.607$
bzip2	344.503	:	$206.584+204.594$
	$C(xy)$	$=$	$C(yx)$
gzip	320.859	:	320.676
bzip2	344.503	:	344.503
	$C(xy)+C(z)$	$\leq$	$C(xz)+C(yz)$
gzip	326.766+192	:	$321.3+309.733$
bzip2	341.933+202.4	:	$340.5+335.866$

5.1.3 同研究室、荒堀、福田よる 4.1.1 のファイルの前処理 2

	$C(xx)$	$=$	$C(x)$
gzip	188.696	:	200.5
bzip2	202.839	:	243.571
	$C(xy)$	$\geq$	$C(x)$
gzip	316.863	:	189.246
bzip2	342.125	:	203.650
	$C(xy)$	$\leq$	$C(x)+C(y)$
gzip	316.863	:	$189.246+188.146$
bzip2	342.125	:	$203.650+202.027$
	$C(xy)$	$=$	$C(yx)$
gzip	316.863	:	316.676
bzip2	342.125	:	342.125
	$C(xy)+C(z)$	$\leq$	$C(xz)+C(yz)$
gzip	315.6+192.8	:	$308.666+314.3$
bzip2	334.366+196.666	:	$333.033+338.533$

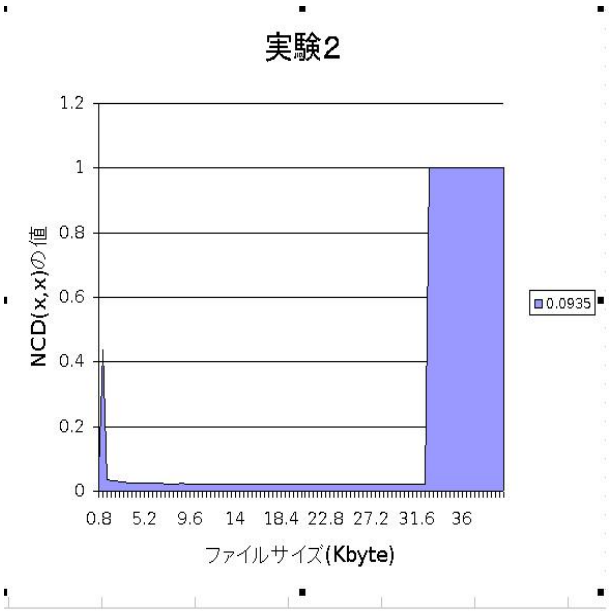
どちらの圧縮アルゴリズムもおおよそ定義を満たしていると言えるだろう。しかし $C(xx) = C(x)$ においては、bzip2 は大きな誤差を生じている。逆に gzip は辞書式の圧縮アルゴリズムなので、同じパターンが文字列中に現れた場合、効率よく圧縮できるので比較的満たしたといえる。つまり方言桃太郎に関して言えば、前処理を行ったファイルを類似度判定で使用するので、圧縮率も高い結果が出た gzip が優れていると言える。

5.2 gzip における NCD(x,x) の限界を調べる

gzip は辞書式のアルゴリズムであるので $C(xx) = C(x)$ を上の実験で比較的満たした。しかし、辞書のサイズには限界があり、ある一点を境に全く機能しなくなるのではないか。そこでファイルサイズを徐々に大きくしていき NCD(x,x) が 0 から乖離するか。類似度判定における gzip の限界を探る。**実験**

使用するファイル：ランダムの文字列からなるファイル (40Kbyte)

実験方法：ファイルの大きさを 400byte ずつ増やしていき、その都度 NCD を計算する。以下が結果のグラフである。



この実験より gzip の辞書サイズは 30kbyte 前後と言える。つまり 30kbyte 以上のファイルに関しては、他の圧縮アルゴリズムを使用することが望ましい。

6 今後の課題

今回の研究では、bzip2 と gzip について、実験を行った。圧縮に基づく類似度判定では ppmz という高圧縮アルゴリズムよく使用されているようである。この ppmz

についても、調べる必要があるだろう。

参考文献

[1] Ming Li, Member, IEEE, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitanyi
The Similarity Metric (2004)

[2] 谷　聖一
谷聖一　データの複雑さ・データ表現の複雑さ

[3] Rudi Cilibrasi and Paul M.B. Vitanyi
Clustering by Compression (2005)