Kolmogorov 記述量を用いた和歌の類似度の算出

Calculation of similar level by Kolmogorov Complexity

谷　研究室　　見目勤
kenmoku tsutomu

概要

日本の古典文学である和歌の中には同じような構文や文脈を用いた類歌という和歌が存在する. これらの和歌の関連性を調べる為, 和歌間の類似度を計る手法として, 新たに Kolmogorov 記述量を用いた類似度の算出方法を提案したい.

1　はじめに

和歌の中には平安時代を中心に「〜なりけり」など特定の構文を用いた和歌が多数存在する. こえっらは類歌と呼ばれ当時の詠歌界において流行った作風の一つである. 古典文学の世界では, 計算機が実用的なレベルで運用可能になったある時期において, これらの類歌の類似性を計算機を使って調べようという試みがあった. しかし, 類歌の種類によって索引を引く事で簡単に見つける事が出来た為に古典文学の世界では急速にその興味が失われていった. その後, 計算機を使った文字列に関する類似度の調査は小説等の長文を扱う国語学においての単語の頻出パターン等の研究で使われるようになる. 今回は, 計算機による新しい類歌判別の手段として,Kolmogorov 記述量を使った距離の測り方から和歌間の類似性を調べ, 単語等によらない新しい類歌のパターンを探したい. 第二章では従来手法について, 第三章では Kolmogorov 記述量について, 第四章では Kolmogorov の実装について, 第五章では実験概要と結果について述べる.

1.1　探索範囲

今回調べる範囲は勅撰和歌集の初代三代集である以下の和歌集を使う事にした.

- 古今和歌集
- 後選和歌集
- 拾遺和歌集

これらの和歌集は平安時代の類歌が盛んな時期にまとめられた為, 沢山の類歌が収録されていると思われる.

2　従来手法

類歌を調べる場合, その和歌の中に含まれている単語や表現等のパーツを抽出し同じパーツを持つ和歌を探す事で類歌を探し出していた.

3　Kolmogorov 記述量

この章では Kolmogorov 記述量の定義、条件付 Kolmogorov,Kolmogorov の距離について説明する.

3.1　Kolmogorov 記述量の定義

Kolmogorov 記述量とはデータの複雑さを表す値であり, 最低限どのくらいの情報量 (プログラムで言えばファイルサイズ) があれば対象を表現出来るかを指す.
一般的に
プログラミング言語 S で記述した x を生成するプログラムのうち最小のサイズ
の事をさす.
s をプログラム言語 $|p|$ をプログラムのサイズとして

$$K_s(x) = \min\{|p| : S(p) = x\} \tag{1}$$

と定義される.
またこの時, プログラム言語はそれが何であれ万能であれば最低限必要な情報量の差は定数分の差しかない事が証明されている.

3.2　条件付 Kolmogorov 記述量の定義

補助情報 y を持つときのデータ x の Kolmogorov 記述量を条件付 Kolmogorov 記述量と呼ぶ.
U をプログラミング言語,$|p|$ をプログラムのサイズとして

$$K_s(x \mid y) = \min\{|p| : S(p, y) = x\} \tag{2}$$

と定義される. 補助情報とはあらかじめ持っている情報で, この情報が生成したいデータの情報を含んでいるとデータを表現するのが簡単になる. もしも, 補助情報が全ての情報を含んでいれば補助情報だけでデータを表現する事が出来る. また, 一般的に補助情報がデータの情報を多少なりとも含んでいれば補助情報が存在する時の Kolmogorov 記述量はデータ単体 Kolmogorov 記述量よりも少くなる. 補助情報 y が持っている x に関する情

報量 I(x:y) を

$$I(x:y) = K(x) - K(x \mid y) \tag{3}$$

と定義される.
I(x:y) と I(y:x) は完全に一致するとは限らないが, 定数分の差を無視すれば対称性を持つ (I(x:y)=I(y:x))
と見なす事が出来る.

3.3 Kolmogorov 記述量の示す距離

距離について考える.
X を集合とする.Xx から非負実数への距離関数 D が X 上の距離であるとは, 任意の X の元 x,y,z に対して, 以下が成り立つ時をいう.
x=y ならば D(x,y)=0
D(x,y)+D(y,z) ≧ D(x,y)(三角不等式)
D(x,y)=D(y,x)
これらについて Kolmogorov の示す距離について考える.Li,chen らの研究では任意の文字列 x,y について以下のような正規化情報距離が提案されている.

$$d(x,y) = \frac{maxK(x \mid y), K(y \mid x)}{maxK(x), K(y)} \tag{4}$$

さらに K(y) ≧ K(x) とすると

$$d(x,y) = 1 - \frac{K(x) - K(x \mid y)}{K(y)} \tag{5}$$

となる.

K(x—y) は K(yx)-K(y) で近似されることが知られているので

$$d(x,y) = \frac{K(yx) - K(x)}{K(y)} \tag{6}$$

となる.
K(xy) は x と y を結合した時の Kolmogorov 記述量. ここでいう距離とは実距離ではなく, 比べる二つのデータがお互いにどのくらいの情報を持っているかを示しており, 距離が近ければ近いほどお互いにデータを持っている. また, 距離が 0 の場合情報が等しいことを意味する.

4 Kolmogorov 記述量の実装

Kolmogorov 記述量の示す距離は計算によってその正確な値を出す事が出来ない為, 圧縮プログラムを使って近似的に値を求める事になる. 尚, 純粋なコルモゴロフ記述量に関しては圧縮率が高ければ高いほど理論値に近づくが, 距離を求める際には平均的な圧縮率が必要になるので実装の際には bzip2 をかける事で近似値を得る.

$$bzip2(x,y) = \frac{bzip2(yx) - bzip2(x)}{bzip2(y)} \tag{7}$$

5 実験

5.1 実験内容

テキストファイルから抜き出した１１１４個の古今和歌集の和歌を全通りで cat で結合し, 結合前のファイル共々圧縮をかけて類似度を計る. 尚, 圧縮を掛ける際に bzip2 ではヘッダファイルが大きすぎて上手く圧縮されなかった為,compress 法で圧縮をかける.

5.2 実験結果

類似度の高かった上位五組
1024kokin 0850kokin 0.351351
1024　恋しきが方もかたこそありと聞け立てれをれどもなき心地かな
0483　片意図をこなたかなたによりかけて会はずを何を玉の緒にせむ

1084kokin 0850kokin 0.351351
1084　美濃の国関の淵河絶えずして君に仕へむよろづ代までに
0850　花よりも人こそあだになりにけれいづれを先に恋ひむとか見し

1024kokin 0798kokin 0.353659
1024　恋しきが方もかたこそありと聞け立てれをれどもなき心地かな
0798　我のみや世を鴬と鳴き詫びむ人の心の花と 散りなば

1024kokin 0482kokin 0.370370
1024　恋しきが方もかたこそありと聞け立てれをれどもなき心地かな
0482　逢ふことは雲居はるかに鳴る神の音に聞きつつ恋ひ渡るかな

1024kokin 0628kokin 0.373494
1024　恋しきが方もかたこそありと聞け立てれをれどもなき心地かな
0628　陸奥に有りと言ふなる名取河無き名取りては苦しかりけり

5.3 考察

上位五組を見た限りでは何とも言えないが, やはり類似度を表す分かりやすい指標のようなものは見つけられない. これがある程度正しい結果なのか, それとも扱って

いるデータや圧縮率の低さが問題で変な結果を出力しているのか, 現状では不確定要素が多すぎるので何とも言えないが, 繋げる順番を逆転させると類似度が大きく変わる辺り cat 結合の段階でゴミか何かが混ざっているのかもしれない. 全体を通してみると類似度の値自体が重なるケースが多く, 類似度の値が 1 を越える組み合わせもいくつか存在した. これは圧縮を書ける前のデータのファイルサイズの上下限がそもそも小さかった事と, 圧縮率が本来予定していた数値よりも低かった事. それらの要因に対して処理するファイルの数が膨大であった事が原因と思われる. 類似度が 1 を越えてしまう組み合わせについては compress を使っても圧縮が出来なかった一部のファイル (約 70 バイト以下) が存在した為, 計算がおかしくなった為だと思われる.

6 今後の課題

テキストに強い圧縮方法の実装

参考文献

[1] 谷聖一　データの複雑さ・データ表現の複雑さ

[2] 三井貴子　KolmogorovComplexity を用いた民族音楽の分類, 2005

[3] 西村久香 系統樹作成アルゴリズムの性能評価実験,2005

[4] 近藤みゆき　古代後期和歌文学の研究

[5] 糸井通告　古代集の文法