4Q-06

木カーネルを用いた文書カーネルの提案とその応用

佐原　諒亮 † 金川　絵利子 ††　岡留　剛 †
† 関西学院大学大学院理工学研究科　†† 関西学院大学理工学部

1　はじめに

作家の文や文章の特徴づけには，文の長さや句読点の間隔・用いる品詞などがよく用いられる．一方,「作家の文体」に焦点をあてた場合，文の句構造や係り受け構造が重要となる．本研究では，係り受け構造に着目して解析を行ない，作家間の文構造の類似度に着目する．文の句構造の類似度を測る指標として木カーネル [1] が存在する．本研究では，木カーネルを用いた文書間の類似度を求める文書カーネルを提案し，それを用いて作家間の類似度を求める．文章中の文の順序も重要な特徴となり得るが，文構造に焦点をあてるために，今回は文の順序は考慮せず bag of sentences に基づいて解析を行なう．

2　木カーネル

木カーネルは，2 つの木構造データ間の共通している構造として，部分木を用いるカーネルであり，共通する部分木の個数を数えることで値が決定される．部分木の定義は以下である．

- 少なくとも 1 個以上の子を持つ任意のノードを選んで，このノード（部分木の根とする）と，子孫ノードの組み合わせで得られる木である．
- 部分木の根以外のあるノードが木に含まれる場合，その兄弟ノードもすべて木に含む．

木カーネルは，木構造 T_1，T_2 に対して，以下の式で定義される．

$$K_A(T_1, T_2) = \langle \boldsymbol{\phi}(T_1), \boldsymbol{\phi}(T_2) \rangle = \sum_{S \in \tau} \phi_S(T_1)\phi_S(T_2),$$

ここで，S は部分木である．τ はすべての固有木の集合で，また，$\phi_S(T)$ は，木 T が S を部分木として含むときは 1，含まないときは 0 となる．これにより，T_1 と T_2 の共通の部分木の数え上げを実現している．

3　文書カーネル

文構造に着目した文書間の類似度を定義するため，本研究では木カーネルを用いた文書カーネルを提案する．2 つの文章中の文同士の木カーネルの平均値を用いているため，本研究で提案するカーネルは「平均文書カーネル」と呼ぶことにする．平均文書カーネルは，1 つ目の文書の各文と 2 つ目の文書の各文との木カーネル値を算出し，算出したすべてのカーネル値の平均を取ることで求める．すなわち，以下が平均文書カーネル $K_D(D_1, D_2)$ の定義である．

1. $K_D(D_1, D_2) = 0, \quad (D_1 = \phi \text{ または } D_2 = \phi)$
2. $K_D(\{s\}, D_2) = \frac{1}{|D_2|} \sum_{s_i \in D_2} k(s, s_i),$
3. $K_D(D_1 \cup \{s\}, D_2)$
   $= \frac{1}{|D_1+1|}(K_D(D_1, D_2) + K_D(\{s\}, D_2)),$

ここで，D_1 は文書 1，D_2 は文書 2，s は文，s_i も文，k は木カーネル，$|D|$ は D 中の文の数を表す．平均文書カーネルは Haussler の論文 [2] より，カーネルであることが証明できる．平均文書カーネルは，1 文と 1 文あたりの共通する部分木の個数の平均であるという意味合いを持つ．

4　評価

平均文書カーネルを用いて実験を行なった．本研究の実験では，青空文庫から比較的作品数の多い 31 作家を選んだ．データのクリーニングのため，係り受け解析器の CaboCha と構文解析器の Mecab を用いた．2 つの文の木カーネル値は葉である単語に大きく依存する．本研究では，構文構造の類似度に焦点を当て，用いられている単語の違いは極力排除したいため，各単語を品詞に還元的に縮約したコーパスを用いる．定義した平均文書カーネルは文数を一致させる必要があるため，各作家の文を 100 文ずつランダムに抽出したコーパスで平均文書カーネルの値を求めた．結果はこれを 10 回行なったものの平均を求めた．また，木カーネルの計算は Moschitti のプログラム [3] を用いて計算した．この際，Subset Trees(SSTs) を用いる方法と，Sub Trees（STs）を用いる方法とがあり，両者について解析を行なった．両者の結果に大きな相違がなったため，今回は STs による解析について述べる．

5　議論

平均文書カーネルによる値を用いて作家間の文体的な特徴による「距離」を算出し，それをもとにバネモデル [4] を作成した（図 1）．バネモデルにおける「距離」は文書カーネルによる値を二乗し，それを逆数に取ったものを用いた．この図からもわかるように，宮沢賢治が他の作家と「距離」が離れており，特徴的な文を用いやすいのではないかと考えることができる．また，芥川龍之介と太宰治の「距離」が近いことから，二人は構文的に似た文を書くのではないか，一方，芥川と夏目漱石は「距離」が離れていることから，二人は構文的には似ていない文を書くのではないかと推測できる．

平均文書カーネルの値が小さい，すなわち，構文的に似ていない文書を書くと思われる宮沢賢治と芥川龍之介に着目する．宮沢は全体的に短い文を書く傾向があり，芥川は宮沢よりも比較的にさまざまな長さの文を書く作家である．1 文が長ければそれに比例して 1

文に含まれる部分木の個数も増加する．一方，1 文が短いとその分 1 文に含まれる部分木は少なくなってしまう．結果として，宮沢と芥川を比較した際，平均文書カーネルの値が小さくなったと言える．次に，芥川と新美を比較する．新美は宮沢と同じく児童文学を書く作家であるが，比較的 1 文の文字数が多めであるため，1 文に含まれる部分木の個数は多くなると考えられる．しかしながら，芥川と新美において平均文書カーネルを算出した場合，値が小さくなった．これは 2 人の作家の構文的な相違が大きく表れた結果であると考えられる．最後に，芥川と太宰を比較する．この 2 人の場合の平均文書カーネルの値は大きい．2 人とも 1 文の文字数は平均して同じ程度であったため，値が大きくなりやすいと考えられるが，新美とは異なり，一致する部分木の個数が多い．また，この 2 人の比較に関しては，宮沢と新美とは異なり，短い文においてもある程度の値を算出していたため，全体として値が大きくなったと言える．

6 関連研究

表層的な文字や記号・品詞の並びに基づく作家の特徴づけに関する研究は多くある．例えば，文の長さについて注目している文献として前川 [5] のものがある．この文献では，各作品の文の長さの平均と，文の長さのばらつきを用いて比較を行なった．単語の長さに関して分析を行なったのが金の研究 [6] である．この文献では，単語の長さの分布には著者の特徴が現れるのか，どうすれば著者の特徴がより明確に得られるかについて分析を行なっている．

読点の打ち方について分析を行なった研究が金らの研究 [7][8] である．ここでは，読点の前の文字や読点の前の文字の品詞，読点を打つ間隔に関する情報の有効性を分析した．いずれにおいても各特徴において作家ごとに特徴がみられると述べている．

英語文書においても同様のような研究が行なわれている．例えば，Yule[9] は，作家の文書の文の長さを分析し，その平均値，中央値，四分位範囲が作家ごとに異なるということを述べた．この結果を受けて Yule は，「The Imitation of Chirst」の著者推定を行ない，文の長さの有用性を示した．

また，テキストデータに対するカーネルとしてベクトル空間カーネルが定義されている．これは単語の出現頻度に基づいたカーネルであり，文の句構造の類似性を反映させる際にはほかの手法を考える必要がある．

7 まとめ

本研究では文書間の類似度を測るため，木カーネルを用いた文書カーネルを提案し，それを用いて評価実験を行なった．実験では係り受け解析を行ない，1 コーパス 100 文と制約を設けたコーパスを用いて実験を試みた．その結果，構文的に似ている作家や，似ていない作家を捉えることができた．文の長さによって結果が左右されやすいが，構文的に似ている作家であれば

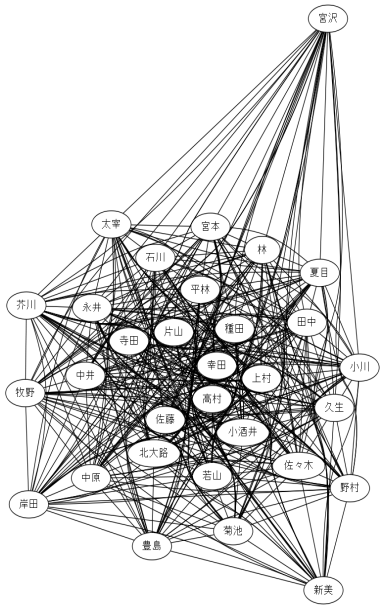

図 1: 文書カーネルの値を用いた作家の文体間の平均「距離」によるバネモデル

平均文書カーネルの値が大きくなりやすいことも確認できた．

参考文献

[1] Collins，M．and N.Duffy (2001)． Convolution kernels for natural language．*Advances in Neural Information Processing Systems 14* [*Neural Information Processing Systems: Natural and Synthetic， NIPS 2001*] ，625-632， MIT Press.

[2] Haussler，D． (1999)．Convolution Kernels on Discrete Structures．Technical report，University of Santa Cruz.

[3] Moschitti，A．TREE KERNELS IN SVM-LIGHT. http://dit.unitn.it/moschitt/.

[4] Kamada，T． and S. Kawai (1989)．An algorithm for drawing general undirected graphs．*Information Processing Letters*，**31**，1，7-15.

[5] 前川守 (1995)．**文章を科学する**，岩波書店．

[6] 金明哲 (1996)． 日本語における単語の長さと分布と文章の著者， **社会情報**， **5**， 2， 13-21.

[7] 金明哲 (1994)． 読点の打ち方と著者の文体特徴，**計量国語学**， **19**， 7， 317-330.

[8] 金明哲， 樺島忠夫， 村上征勝 (1993)． 読点と書き手の個性， **計量国語学**， **18**， 8， 382-391.

[9] Yule，G．U．(1939)． On sentence-length as a atatistical characterristic of style in prose; with application to two cases of disputed authorship，*Biometrika*， **30**，3，363-390.