数理解析研究所講究録
1489巻 2006年 209-215

部分木のマッチを用いた構文解析木間の類似度について

秋友　克俊(Katsutoshi Akitomo)
岡山理科大学大学院(Okayama University of Science Graduate College)

椎名　広光(Hiromitsu Shiina)
岡山理科大学(Okayama University of Science)

1. はじめに

新聞記事や本などから作られた電子化された文情報の中から欲しい文をコンピュータの検索機能を利用して目的の文を調べている. そのとき検索を行ったとき目的とされることが多い文が上位に出力されるように作られている[2][3][5]が, 目的以外にも多くの結果が出力されてしまう. その原因として単語には複数の意味が存在しているため, 検索語に指定した単語が最もよく使われる意味で使われている文が検索結果の上位に出力されることが挙げられる. よって検索語にした単語があまり使われない意味で用いられている文を調べたい場合に, 目的の文やページが下位に出力されるため出力の中からさらに探す必要がある. このような問題は, 検索語の意味を特定する精度を良くして検索語を異なった意味で使用している目的以外の文を減らすことで解決することができると考えられる. その検索語の意味を特定する方法として, 数個の単語からなる短い文を利用して検索したい単語とその前後の単語の品詞の関係を調べる方法を想定し, その基礎研究として文の中で使われている単語間の品詞の関係である構文解析木[4][6]の類似度を調べることで検索語と同じ意味で検索語が使われている文を調べることができるか調査する.

本研究で利用する構文解析木の特徴は, 文の中で使われている単語間の品詞的な関係を木構造で示していることである. 従って, 構文解析木の類似度は, 文同士の構文解析木の構造の一致具合を数値にしたものであり, 単語が同じ意味で使われている場合には両方の文で同じ構造が存在するため類似度は高くなる. また, 文中の調べたい単語以外の単語の品詞構造が比較した文の構造と一致して類似度が高まることもある. そこで, 文間の構文解析木の類似度をいくつかの方法で調べ, 文の内容を最も反映することが出来る類似度計算方法を調べる. その類似度計算方法として構文解析木の一致具合を優先した計算法や構文解析木の大きさを優先した計算方法を提案する. また, 構文解析木は接続詞などの小さな差で変化してしまうので, 変化した部分を置き換えて変化する前の構文解析木との類似性を保つ方法として部分木の一部を入れ替えることで部分木の拡張を行う方法について調べる.

2. 諸定義

本論文で使用する記号の準備として, 文法をG=(N,T,P,S) N:非終端記号 T:終端記号 P:生成規則 S:初期記号で表し, この文法Gを用いて文AとBから生成される構文解析木をT_AとT_Bとする. 部分木を構文解析木から抜き出された木構造の一部と定義し, 構文解析木をT_Aに対してその部分木をαと表わし, その部分木の個数をN_Aとする. ここで本論文で使用する文法規則を図 1に示す.

A→B	B→C A	B→D E	C→E	E→C C
A→B A	B→C D	B→D H	C→F	F→D E
A→B C	B→C E	B→F G	C→F G	

図1:文法規則の一覧

次に本論文で使用する構文解析木から抽出される部分木の集合を定義する.

① 構文解析木T_Aの部分木の集合$SF(T_A)$

構文解析木T_Aの部分木の集合を$SF(T_A)=\{\alpha_1,\alpha_2,\ldots,\alpha_{N_A}\}$と表現する. 図2に構文解析木$T_A$, 図3に図2の構文解析木から抽出される部分木列$SF(T_A)$を示す.

図2:構文解析木T_A

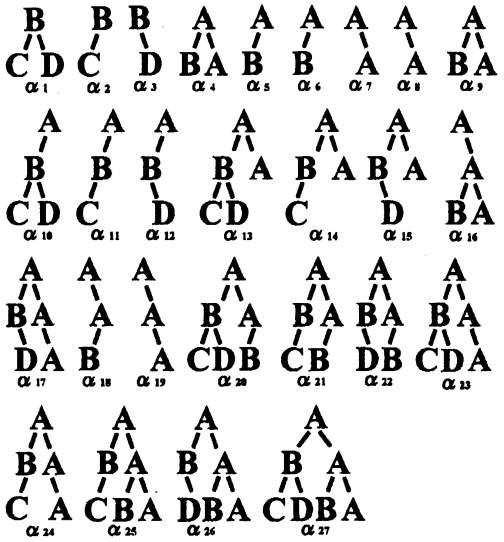

図3:構文解析木T_Aの部分木列$SF(T_A)$

② 部分木列$SF(T_A)$から重複を取り除いた部分木列$SFA(T_A)$

抽出した部分木列$SF(T_A)$から重複する部分木を取り除いた部分木列を$SFA(T_A)$とする. $SF(T_A)$を用いて類似度を求めるとき, 部分木が重複して存在すると, 一致個数の最大値は互いの文の部分木の個数の積で求めるしかない. これでは部分木の個数の積を一致個数の最大値とした比較方法しか作ることが出来ない. そこで重複する部分木を一つにまとめた部分木列$SFA(T_A)$を提案する. 図4に$SF(T_A)$から作られる$SFA(T_A)$を示す.

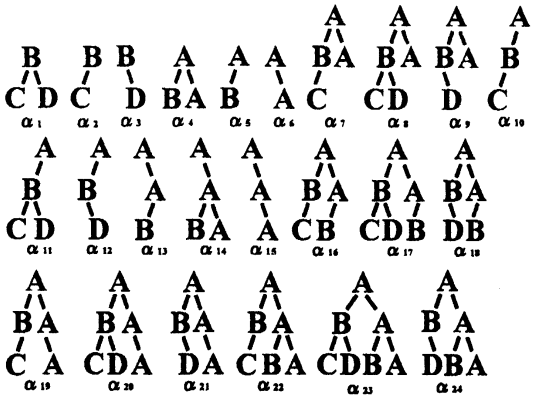

図4:重複を取り除いた部分木列$SFA(T_A)$

③ 部分木を拡張した部分木列$SF(T_A)^+$

文の部分木の葉に他の文法規則から生成される部分木を付加することで部分木の一部を入れ替えて部分木の拡張を定義する. 部分木の拡張の例を図5に示す.

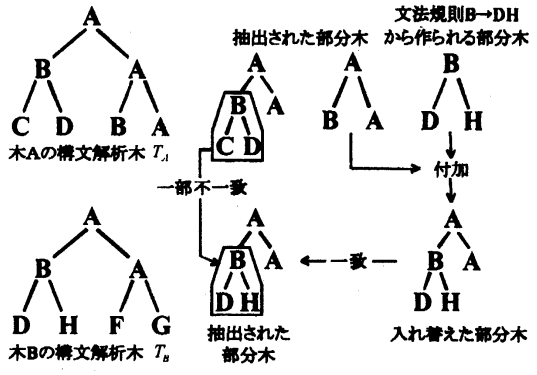

図5:部分木の拡張の例

ここで部分木列$SF(T_A)$の部分木を文法規則から作られる木を用いて拡張した部分木列ADD(部分木列,文法規則から作られる部分木)を定義し, これを$SF(T_A)^+$と表す. また, 拡張した部分木を加えた部分木の個数をN_A^+とする.

$$ADD(SF(T_A),P)=\{\alpha_1,\alpha_2,\ldots,\alpha_{N_A^+}\}$$

$$SF(T_A)^+ = ADD(SF(T_A),P)$$

3. 構文解析木の類似度

本章では構文解析木の部分木の比較を用いた類似度の求め方について説明する. 本論文の類似度は二つの文の部分木列から部分木の一致個数を調べ, 一致個数と比較回数や部分木の個数

を用いて求める．また，構文解析木内での位置関係や木の大きさを考慮に入れて求める方法も提案する．なお，構文解析木は文に対して一意に求められないことがあるが，本研究では一つの文に構文解析木が一つのみ存在するとする．

3.1 部分木の一致の計算方法

部分木の一致個数計算として部分木の一致と高さ情報を考慮した計算について説明する．

① 部分木の一致個数計算法

文の類似度を調べるために2つの文の構文解析木に含まれる部分木の部分木集合を比較し一致個数を求める．部分木の一致個数を計算するために，部分木列 $\alpha=\{\alpha_1,\alpha_2,\ldots,\alpha_{N_\alpha}\}$，$\beta=\{\beta_1,\beta_2,\ldots,\beta_{N_\beta}\}$ に対して，sum_1 を以下のように定義する．ここで一致個数を sum_1，部分木 α と β の比較は $com(\alpha_k,\beta_l)$ で表す．sum_1 を式で表すと次のようになる．

$$sum_1(\alpha,\beta)=\sum_{k=1}^{N_\alpha}\sum_{l=1}^{N_\beta}com(\alpha_k,\beta_l)$$

なお，部分木の比較の式 com は以下のとおりである．

$$com(\alpha_k,\beta_l)=\begin{cases}1 & \alpha_k=\beta_l\text{のとき}\\ 0 & \alpha_k\neq\beta_l\text{のとき}\end{cases}$$

② 部分木の高さを使った計算法

部分木は高さが大きいほど多くの情報を持ち，抽出される前に葉に近いところに存在していたものほど文の特徴を現しやすい．そこで，部分木の高さと抽出前の場所を定義して類似度の計算に加える．部分木 α_n の高さを $H(\alpha_n)$ とし，抽出する木の中で α_n が存在していた高さを $O(\alpha_n)$ とする．また，要素の次数を $R(\alpha_n)$ とする．図6に構文解析木の要素の次数について示す．

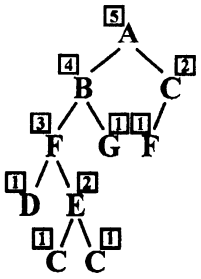

図6：構文解析木の要素の次数 $R(\alpha_n)$ の例

$O(\alpha_n)$ は次の式で求められる．

$$O(\alpha_n)=R(\alpha_n)-H(\alpha_n)+2$$

部分木列の一致個数を $O(\alpha_n)$ と $H(\alpha_n)$ を用いて補正した計算 sum_2 を示す．

$$sum_2(\alpha,\beta)=\sum_{k=1}^{N_\alpha}\sum_{l=1}^{N_\beta}\frac{H(\alpha_k)\cdot com(\alpha_k,\beta_l)}{O(\alpha_k)+O(\beta_l)}$$

3.2 部分木の類似度

本論文では類似度に部分木の一致の計算法，部分木の重複，部分木の高さ情報を組み合わせて計算している．表1に 3.1 で示した sum_1 を利用した類似度計算，表2に sum_2 を利用した類似度計算を示す．以下では sim_1 ～ sim_{10} の定義を示す．

表1：sum_1 を利用した類似度計算の一覧

	評価の指針	部分木の拡張なし	部分木の拡張あり
部分木の重複あり	一致個数を優先	sim_1	sim_2
部分木の重複なし	一致個数を優先	sim_3	sim_4
	部分木が少ない方を優先	sim_5	sim_6
	相互の文の類似度の平均	sim_7	sim_8

表2：sum_2 を利用した類似度計算の一覧

	評価の指針	部分木の拡張なし	部分木の拡張あり
部分木の重複あり	一致個数を優先	sim_9	sim_{10}

①部分木列 $SF(T_A)$ と部分木列 $SF(T_B)$ の一致個数 sum_1 を用いた部分木の一致個数を優先した類似度の式 sim_1 を次のように定義する．

$$sim_1(T_A,T_B)=\frac{sum_1(SF(T_A),SF(T_B))}{N_A\cdot N_B}$$

②部分木を拡張した部分木列$SF(T_A)^+$と部分木列$SF(T_B)$を用いた類似度sim_2を定義する.

$$sim_2(T_A,T_B)=\frac{sum_1(SF(T_A)^+,SF(T_B))}{N_A^+\cdot N_B}$$

③部分木列$SFA(T_A)$と部分木列$SFA(T_B)$を用いた類似度の式sim_3を定義する.

$$sim_3(T_A,T_B)=\frac{sum_1(SFA(T_A),SFA(T_B))}{N_A\cdot N_B}$$

④部分木列$SFA(T_A)^+$と部分木列$SFA(T_B)$を用いて計算した式sim_4を定義する.

$$sim_4(T_A,T_B)=\frac{sum_1(SFA(T_A)^+,SFA(T_B))}{N_A^+\cdot N_B}$$

⑤部分木列$SFA(T_A)$と部分木列$SFA(T_B)$の一致個数の最大値は部分木の個数N_AとN_Bの小さい方と同じになる. そこで, 部分木が少ない文を優先した類似度の式をsim_5を定義する.

$$sim_5(T_A,T_B)=\begin{cases}\dfrac{sum_1(SFA(T_A),SFA(T_B))}{N_A} & (N_A\le N_B\text{のとき})\\ \dfrac{sum_1(SFA(T_A),SFA(T_B))}{N_B} & (N_A>N_B\text{のとき})\end{cases}$$

⑥部分木列$SFA(T_A)^+$と部分木列$SFA(T_B)$の部分木が少ない文を優先した類似度の式をsim_6に定義する.

$$sim_6(T_A,T_B)=\begin{cases}\dfrac{sum_1(SFA(T_A)^+,SFA(T_B))}{N_A^+} & (N_A^+\le N_B\text{のとき})\\ \dfrac{sum_1(SFA(T_A)^+,SFA(T_B))}{N_B} & (N_A^+>N_B\text{のとき})\end{cases}$$

⑦部分木の一致個数は比較する文が長い文であるほどに大きくなる傾向がある. そこで, 部分木列$SFA(T_A)$と部分木列$SFA(T_B)$と部分木の個数N_AとN_Bを用いて類似度sim_7を定義する.

$$sim_7(T_A,T_B)=\frac{1}{2}\left\{\frac{sum_1(SFA(T_A),SFA(T_B))}{N_A}+\frac{sum_1(SFA(T_A),SFA(T_B))}{N_B}\right\}$$

⑧部分木列$SFA(T_A)^+$と部分木列$SFA(T_B)$の部分木の個数N_A^+とN_Bを用いて類似度sim_8を定義する.

$$sim_8(T_A,T_B)=\frac{1}{2}\left\{\frac{sum_1(SFA(T_A)^+,SFA(T_B))}{N_A^+}+\frac{sum_1(SFA(T_A)^+,SFA(T_B))}{N_B}\right\}$$

⑨部分木列$SF(T_A)$と部分木列$SF(T_B)$の一致個数を$O(\alpha_n)$と$H(\alpha_n)$を用いて補正した値sum_2を用いた部分木の一致個数を優先した類似度の式sim_9を定義する.

$$sim_9(T_A,T_B)=\frac{sum_2(SF(T_A),SF(T_B))}{N_A\cdot N_B}$$

⑩部分木列$SF(T_A)^+$と部分木列$SF(T_B)$の一致個数を$O(\alpha_n)$と$H(\alpha_n)$を用いて補正した値sum_2を用いた部分木の一致個数を優先した類似度の式sim_{10}を定義する.

$$sim_{10}(T_A,T_B)=\frac{sum_2(SF(T_A)^+,SF(T_B))}{N_A^+\cdot N_B}$$

3.3 類似度の調査例

構文解析木の例T_A, T_B, T_Cを用いて構文解析木から部分木を抽出し類似度を求める例を示す.

① T_A, T_B, T_Cから抽出できる部分木の類似度

構文解析木からの部分木抽出の例を図7に示し, 一致した部分木を図8に示す.

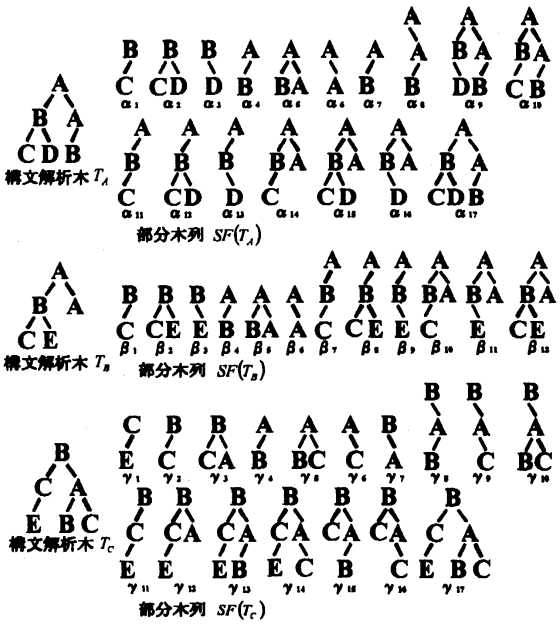

図7: T_A, T_B, T_Cの部分木抽出

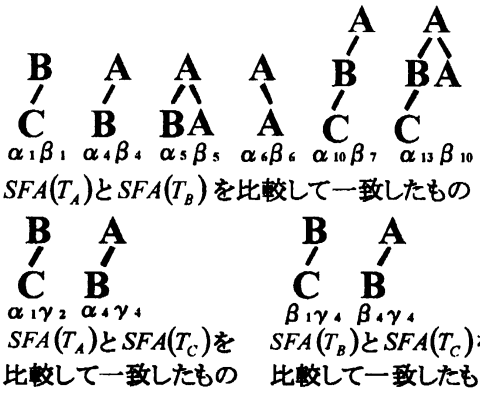

図8：T_A，T_B，T_C 間の一致した部分木

図8の一致結果から部分木の存在した高さと部分木の高さを調べ図9に示す．

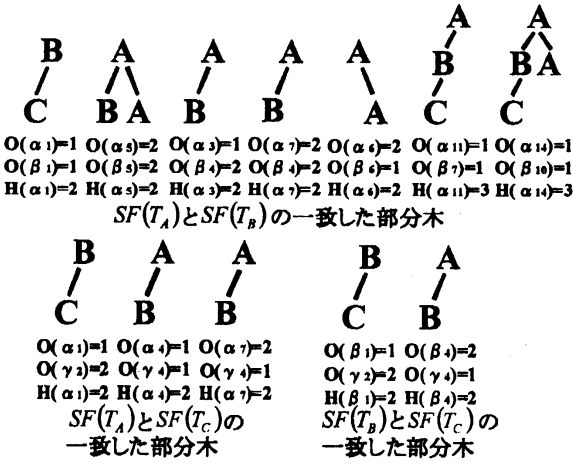

図9：比較結果の $O(\alpha_n)$ と $H(\alpha_n)$

T_A，T_B，T_C に含まれる部分木の個数は $N_A=17$，$N_B=12$，$N_C=17$ である．部分木の個数と一致個数から求められる類似度 sim_1 と sim_9 の計算結果を表3に示す．

表3：類似度 sim_1 と sim_9

	T_A, T_B	T_A, T_C	T_B, T_C
sim_1	0.0343	0.0104	0.0098
sim_9	0.0310	0.0081	0.0065

② 重複を取り除いた部分木列の類似度

図8で示した部分木列から重複を取り除いた部分木列を図10に，図10の部分木の集合から一致する部分木を調べた結果を図11に示す．

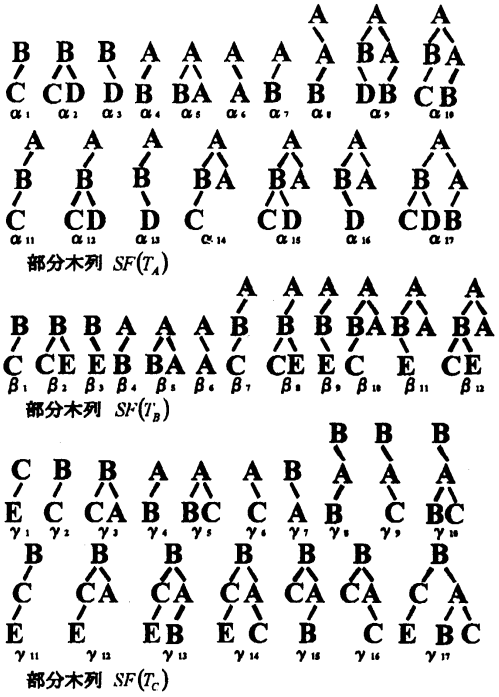

図10：重複を取り除いた部分木列

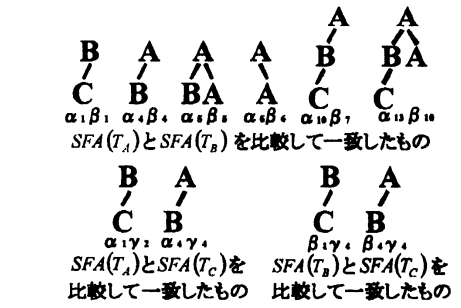

図11：重複を取り除いた部分木の一致

部分木の一致個数と部分木の個数 $N_A=16$，$N_B=12$，$N_C=17$ を用いた類似度の計算結果を表4に示す．

表4：類似度 sim_3，sim_5，sim_7

	T_A, T_B	T_A, T_C	T_B, T_C
sim_3	0.0313	0.0074	0.0098
sim_5	0.5000	0.1250	0.1667
sim_7	0.4375	0.2426	0.2843

③ 部分木の拡張

図1で定義した文法規則に従って図8で示した部分木列 $SF(T_A)$ の部分木の拡張し $SF(T_A)^+$ を作成する．部分木の拡張の例を図12に，拡張した部分木列 $SF(T_A)^+$ の例を図13に示す．

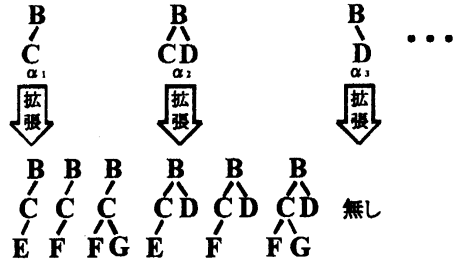

図12:部分木列$SF(T_A)$の部分木の拡張例

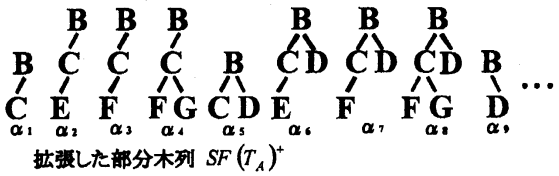

図13:拡張した部分木列$SF(T_A)^+$の例

部分木列の部分木の個数$N_A=155$，$N_B=133$，$N_C=211$と重複を取り除いた部分木列の部分木の個数を$N_A=148$，$N_B=133$，$N_C=211$を用いた類似度の計算結果を表5に示す.

表5:部分木を拡張した場合の類似度

	T_A, T_B	T_A, T_C	T_B, T_C
sim_2	0.0031	0.0006	0.0004
sim_4	0.0029	0.0004	0.0004
sim_6	0.4286	0.0902	0.0827
sim_8	0.4069	0.0690	0.0674
sim_{10}	0.0048	0.0008	0.0005

4. 調査実験

本研究では EDR 電子化辞書の日本語コーパス[1]に収録されている構文解析木を用いた. この中の 1000 個の構文解析木から部分木を抽出し, 文 1000 個の中で使用されている全ての文法規則 577 個を使用して部分木を拡張し調査した.

① 評価法ごとの順位の変化

調査の結果, 部分木を拡張と部分木の高さと抽出前の場所を用いた拡張による順位変動を示す. ここで図14にsim_1, sim_2, sim_3, sim_4, sim_9, sim_{10}における文番号 1:JCO0000217 の順位変動のグラフを示す.

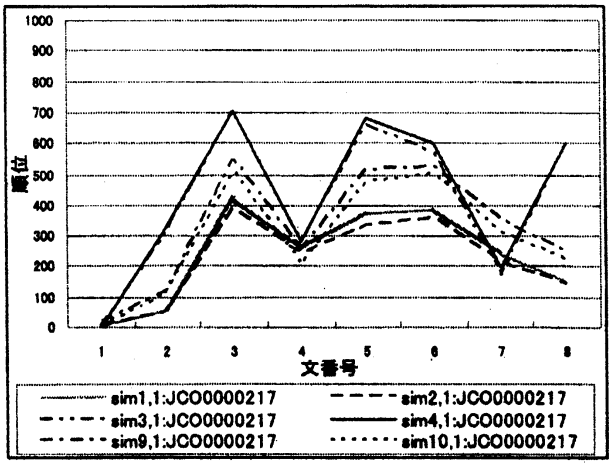

図14:評価方法ごとの順位変動

この結果から部分木の拡張による大きな順位の変動が無いことが分かる. また部分木の重複をなくすことで順位が変動しやすくなり, 部分木の高さと抽出前の場所を用いた拡張では極端に順位が変動しやすくなっている.

②順位の偏りの変化

部分木の重複をなくしたときの調査では, sim_7を用いたときに上位に来るものの偏りが比較的少なかったので, 今回の調査では sim_7 の評価法が良いと考えられる. ここで各評価方法における構文解析木が極端に小さく出現頻度の高い部分木で構成されているために偏りやすい文 556:JCO0010899 の順位変動を図15に示す.

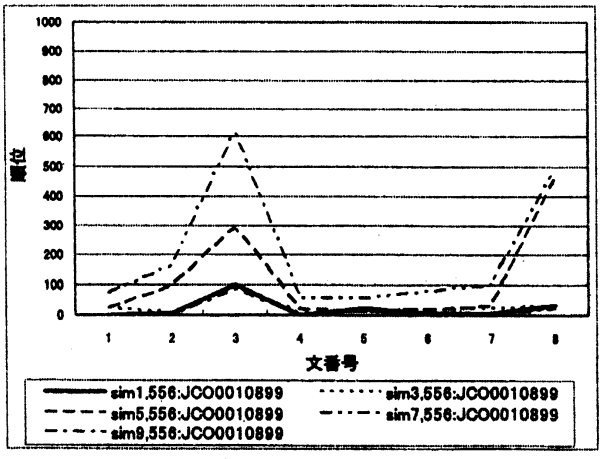

図15:評価方法ごとの順位変動

この表からsim_5，sim_7を用いたときに構文解析木が極端に小さい場合でも類似度を変化させることが出来ると分かる．

③類似度計算法ごとの順位の差

類似度の各計算方法によってその結果がどれぐらい似ているかを調査した．調査方法は，sim_1からsim_{10}の10通りの類似度計算方法に対して1:JCO0000217から50:JCO0001050までの50文に，1:JCO0000217から1000:JCO0021479までの1000文との類似度を調べ，1位から1000位の順位を付けた．その順位の上位20位までの順位の差を合計した値を用いて調査した．表6に結果を示す．

表6:部分木を拡張していない場合の
類似度調査法ごとの順位変動の例

	sim_1	sim_3	sim_5	sim_7
sim_1	—	—	—	—
sim_3	12,418	—	—	—
sim_5	15,892	9,746	—	—
sim_7	15,614	12,632	12,060	—
sim_9	11,890	14,138	16,722	16,692

表6の結果から，重複をなくした場合の類似度計算法sim_4 sim_6 sim_8は互いによく似た結果になることが分かる．特にsim_4とsim_6はほとんど同じ結果が求められるため両方の必要性は低いと考えられる．

6. まとめ

類似度を調査するために最も適した計算方法を調べるために，文から抽出した部分木に対して2種類の一致個数計算法を10種類の類似度計算法を示した．その中では，部分木の一部を入れ替えることで部分木を拡張することが出来，よく似た構造を持つ部分木も類似度に反映することができた．また，部分木を構文解析木から抽出する前の場所と部分木の大きさを用いての一致個数の補正は，補正を行わなかった場合に比べて文の特徴を反映した類似度を求めることが出来たと考えられる．

10種類の類似度計算方法によって類似度を数値として求めることができることは分かったが，有効性を調べるためには今後部分木を抽出した文と部分木の類似度の関係を調べる必要がある．また今後の類似度調査法に関する課題は，部分木の重複を取り除いた部分木列に対して部分木を構文解析木から抽出する前の場所と部分木の大きさを用いての補正を行うことである．

また，部分木の抽出や比較に使ったアルゴリズムについて，部分木の抽出はあまり無駄の無いアルゴリズムで十分な速度がある．比較は2つの文の部分木を互いに全て比較を行っているのが，抽出した部分木を根で並べ替えるなどして1つの部分木ごとに根が同じ部分木のみ比較するなど高速化が望めると思われる．

参考文献

[1] EDR 電子化辞書，㈱日本電子化辞書研究所，1998.

[2] 言語情報処理，岩波書店，長尾真，黒橋禎夫，佐藤理史，池原悟，中野洋，1998.

[3] 情報検索と言語処理，東京大学出版，徳永健伸，1999.

[4] 自然言語処理，岩波書店，長尾真，佐藤理史，黒橋禎夫，角田達彦，1996.

[5] 情報検索アルゴリズム，共立出版，北研二，津田和彦，獅々堀正幹，2002.

[6] 確率的言語モデル，東京大学出版社，北研二，1999.

[7]単語と頻度統計を用いた類似性の定量化，電子情報通信学会論文誌，Vol.J87-DII No.2，pp.611-672，深谷亮，山村毅，工藤博章，松本哲也，竹内義則，大西昇，2004.

[8] アルゴリズムとデータ構造，コロナ社，湯田幸八，伊原充博，2001.