自然言語処理　150－15
(2002. 7. 16)

極大類比による文書群の組織化

原口　誠†　中野　卯虎†　吉岡　真治†

† 北海道大学大学院工学研究科電子情報工学専攻　〒060-8628 札幌市北区北１３条西８丁目
E-mail: †{makoto,yoshioka}@db-ei.eng.hokudai.ac.jp

あらまし　類似した文書群に共通するイベント系列により文書群を組織化する可能性について論じる。そのために、文の格構造の汎化を要素としてもつ極大類比を導入し、汎化コスト条件を用いた極大類比の発見的検出手法を示す。あわせて、非類似文書とのネガティブマッチングにより、可能な類比を抑制する効率の良い方法も示す。意味処理の困難さに起因する問題点についても触れ、それに関する今後の計画も述べる。

キーワード　極大類比

Organization of Documents in Terms of Maximal Analogies

Makoto HARAGUCHI†, Shigetora NAKANO†, and Masaharu YOSHIOKA†

† Div. Electronics and Information Eingineering, Graduate School of Engineering, Hokkaido University, N13, W8, Kita-ku, Sapporo 060-8628, Japan
E-mail: †{makoto,yoshioka}@db-ei.eng.hokudai.ac.jp

Abstract　Given two or more similar documents in the form of texts, we present a notion of maximal analogies representing maximal sequences consisting of pairs of similar events in the documents. They are required to satisfy certain cost condition so that meaningless similarities between documents are never concluded. A bottom-up search procedure to find a maximal analogy satisfying the cost condition is also presented. In addition to the set of similar documents, we suppose another set of dissimilar ones. Then a maximal analogy is furthermore tested for their appropriateness so as not to explain the latter ones. The test can be performed by an effective subsumption check procedure.

Key words　Maximal Analogy

1. 問題の所在と形式化の背景

電子化された多様な文書群への高速で柔軟なアクセスを目指して、文書の検索、マイニング、クラスタリング、要約等の様々な技法が提案されている。対象とするデータ量が膨大であるために、文書を語彙集合とみなす立場、すなわち、キーワードや索引語による方法論に限定するか、もしくは、タグを導入した(半) 構造化文書を対象にすることが多いと思われる。キーワードは処理の容易さのわりには、実用上許容できる程度の文書の識別・被覆能力を持つために、極端に困るということはないが、表現力の欠如に起因する精度落ちは悩みの種である。一方、タグ付けされた文書は高度な処理が可能となるが、タグ付けに要するコストは無視できるものではない。

こうした問題状況に鑑み、本研究では、キーワード集合で表現された文書とタグ付け文書の中間物として位置付けできる文書表現に基づいて、文書群を組織化し、多様な立場・観点から文書へアクセスできる手法の開発を目指している。この目的のために、組織化手法に対し下記の性質を要請する：

R1　文書が持つ物語性を反映できる「索引付け」が基本的には自動でできること、

R2　文書への様々なアクセスを保障するために、できるだけ多様な「索引」を持てること、

R3　個人もしくは集団によって異なる索引付けができること、

R4　質問・アクセス要求に対して、関連する文書を容易に特定できること

ここで、『誰が何に対してどのようなことを行った』かというイベントとそれらの間の関係は物語性を特定するために不可欠である。一般にイベントのフローは、時系列としての出来事の並列性も含んでいるが、本研究では簡単のために文書の索引としてイベント列のみを考察する。並列構造は、複数のイベント列解析によって抽出可能であり、後処理として検出することを予定している。こうしたイベント列による索引付けは、キーワードによる索引付けと同様に、重要なイベント列のみが考慮されるべきである。情報検索や文書要約で用いられている重要

度に関する様々な経験則、例えば TF/IDF 等を用いるのが標準的だと思われるが、経験則の適用によって落ちてしまうイベント列もできるだけ保持したい。それゆえに、本研究では、複数文書間の文書要約手法 [5] とのアナロジーから、下記の仮定と要請をおく：

> 特定の文脈に照らして重要なイベント列は、個人・集団がある観点から類似しているとみなす複数の文書に共通に含まれる。逆に、そうした文書群に共通に出現するイベント列は重要なものである可能性があり、そうした全ての可能性をフォローできること。

この仮定と要請のもとに、

> 類似文書群に共有される類似したイベント列を汎化したイベント列を索引として定める。汎化は各イベントの記述要素たる語彙の汎化とイベントそれ自体の係り受け・格構造の構造汎化からなる。

汎化イベント列も一つのイベント列であることから、一つの文書を表現しており、類似した文書から共通な部分を別の文書として抽出したと理解してよい。また、イベント列に対して、語彙と構造の包摂関係に基づいた自然な順序・包摂関係を導入できる。この包摂関係の下で、汎化イベント列も含めて、全てのイベント列がなす空間は順序集合としての構造を有する。この単純な事実は、質問・検索言語を設計するときの指針を与える。すなわち、質問式も一つのイベント列として与え、より特殊な索引イベント列を持つ文書を関連文書と定めることができる。これは、キーワード集合で文献を表現したときの集合の包含関係による関連性の定義を、イベント列に対して拡張したものになっている。本稿では、実際の質問言語の設計は行ってはいないが、イベント列とその汎化を考察する重要な根拠をなしているので、あえてここで指摘しておく。

文書中の各文に対して、語彙と構造の抽出のために、形態素解析および構文解析器を用い、語彙と係り受け・格関係からなる概念グラフ [2] を導出し、一つのイベントを表すとする。意味解析は重要となるが、現時点では組み込んでおらず、動詞や複合名詞の標準化処理のみを行っているに過ぎない。今後、どの程度の意味処理が必要となるかは、対象とする文書の質・量および、キーワードの拡張たるイベント列の品質レベルとしてどこまで要求するかに依存する。キーワードのみで実用上は使える実態を考えれば、その拡張たるイベント列に対してもおのずから適切な要求レベルが決まってくると期待している。これらの問題に関する現状と課題の考察は、最後の節 6. でまとめておく。

2. 前処理による概念グラフの抽出

形態素解析および構文解析結果から、ノードに品詞を、リンクに係り受け・格でラベル付けした木構造が導出される。本研究では、木構造を概念グラフ [2] と捉え、概念グラフ間の包摂関係を木構造に特化した包摂関係を用いる。木構造の汎化は、パスの捨象とノードのラベルたる語の汎化からなる。語の汎化のために、EDR 電子化辞書 [3] を用いている。すなわち、(単語辞書中の) 単語および (概念辞書中の) 概念ＩＤの集合 $Term$ 上の半順序 $\preceq$ を下記で定めておく。簡単のため、$Term$ の要素を単に項と呼ぶ。

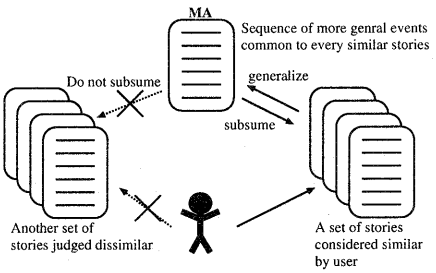

図 1　拡張インデックスとしての極大類比

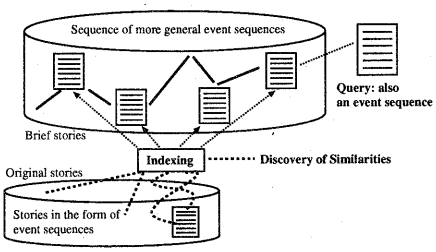

図 2　拡張インデックスによる文書群の組織化

$t_1 \prec t_2$ iff　(1) t_1, t_2 はともに概念ＩＤで、t_2 は t_1 の上位概念である、または、

(2) t_1 は単語辞書中の語、t_2 は概念辞書中の概念ＩＤで、t_1 に関連づけられたある概念ＩＤ t_3 で、t_2 よりも概念辞書において下位のものが存在する。

以下、『誰が何をどうした』という情報単位を、イベントと呼び、節点に $Term$ 中の項をラベルに、また、エッジに格もしくはロール記号を付与した概念木で表記する。特に、動詞中心の整理を行うために、木の根は文の動詞にとっておく。

[Definition 1]　prefix-complete なラベル列 (パス) の集合 $Path(g)$、ならびに、パスに対する項の割り当て $term_g$ の組 $(Path(g), term_g)$ でイベント g を表現する。２つのイベント $g_j = (Path(g_j), term_{g_j})$ に対し、$Path(g_2) \subseteq Path(g_1)$ かつ $term_{g_1}(p) \preceq term_{g_2}(p)$ が全ての $p \in Path(g_2)$ で成り立つとき、g_2 は g_1 を包摂する、もしくは、より一般的であると言い、$g_1 \preceq g_2$ と記す。 □

g_1 の汎化である g_2 は、g_1 中の一部のパスを除去し、残ったパスに割り当てられた項をより一般的な項に置き換えて構成できることを意味している。

3. 極大類比

イベント間の包摂関係が定まると、最小汎化に対応する極小汎化 $MCS(g_1, g_2)$ がイベント g_j 間の類似性表現として求まる。

$$MCS(g_1, g_2) = (Path(g_1) \cap Path(g_2), mst).$$

ただし、mst は、g_1,g_2 の共通パス p が持つ項 $term_{g_1}(p)$, $term_{g_2}(p)$ の極小上位語（極小上界）を選択する関数であり、一般には複数個存在する。図 3 にMCSを例示しておく。図において、$t = \{t_1, ..., t_n\}$ なる表記は、項 $t_1, ..., t_n$ の極小上界 t を mst の値として選択したことを示す。また、特に、$p \in Path(g_1) \cap Path(g_2)$ のラベルとして、$mst(p)$ のかわりに、2つの項からなる集合 $\{term_{g_1}(p), term_{g_2}(p)\}$ でラベルづけしたものをMCSのスケルトンと呼び、$SCS(g_1, g_2)$ と記す。

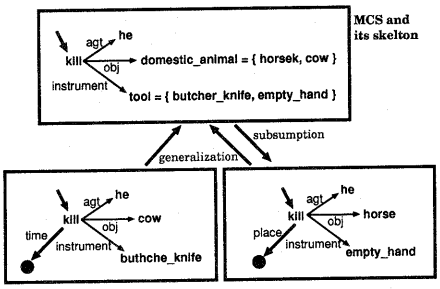

図 3　MCSとそのSCS

このように、2つの文書中に現れる2つのイベント対に対して、その極小汎化として類似性を捉えるのは容易である。問題は、複数のイベント対を一斉に汎化して得られる類似性をいかに定めるかである。本研究ではイベントの対応関係で、イベントの出現順序を保存するイベント対の族のみを考察する。これは、節 1. で述べたような、出来事の時系列としての類似性に着目することを意味する。

[Definition 2]　文書 D_j をイベント $g_k^{(j)}$ の列 $D_j = g_1^{(j)}, ..., g_{n_j}^{(j)}$ として与え、$g_i^{(j)} < g_{i+1}^{(j)}$ なる全順序を仮定する。このとき、D_j のイベントの組の集合 $\theta = \{< g_{n_1}^{(1)}, g_{m_1}^{(2)} >, ..., < g_{n_\ell}^{(1)}, g_{m_\ell}^{(2)} >\}$ で順序を保存するもの、すなわち、$g_{n_i}^{(1)} < g_{n_j}^{(1)}$ ならば $g_{m_i}^{(2)} < g_{m_j}^{(2)}$ となるものを op-選択と呼ぶ。 □

各 op-選択 θ に対して、項の汎化方法がイベント列の汎化を構成する個々のイベント対の汎化に依存せずに、列全体で決まることを要請する：

> **整合性条件**：イベント列において、同一の項が複数の上位語に汎化されることを禁止する。

形式的には、op-選択 $\theta = \{< g_{n_1}^{(1)}, g_{m_1}^{(2)} >, ..., < g_{n_\ell}^{(1)}, g_{m_\ell}^{(2)} >\}$ の各イベント対に対して定まる極小汎化 $MCS(g_{n_j}^{(1)}, g_{m_j}^{(2)})$ の共通パス p に付随した項 $term_{g_{n_j}^{(1)}}(p), term_{g_{m_j}^{(2)}}(p)$ を同一視できる最小の同値関係 $\sim_\theta$ に対する制約であり、

> $\sim_\theta$ に関する同値類に属する語 w_1, w_2 に対し、同一の汎化項を割り当てる

を意味している。

次に、各同値類 $[w]$ 毎に一つの極小汎化語を選択する関数 mst に対して、イベント対から構成される極小汎化 $MCS(g_{n_j}^{(1)}, g_{m_j}^{(2)})$ 中の項割り当てを $mst(p)$ に一斉に置き換えてできるMCS列

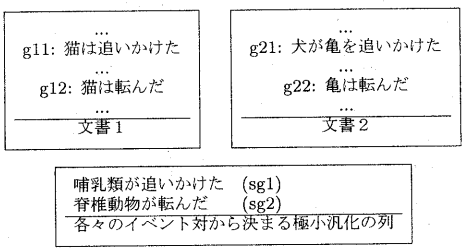

図 4　op-選択 {<g11,g21>,<g12,g22>} に対する極小汎化列と整合性条件

$MCS(g_{n_1}^{(1)}, g_{m_1}^{(2)}), ..., MCS(g_{n_\ell}^{(1)}, g_{m_\ell}^{(2)})$ を $MCCS(\theta, mst)$ と記す。例えば、op-選択の最初の対から形成された極小汎化が、「猫」と「犬」を「哺乳類」に、2つめの対からは「猫」と「亀」を「脊椎動物」に汎化したとする（図 3.）。この2つの極小汎化の単純な列挙では、「猫」は「脊椎動物」と「哺乳類」に多重汎化され、列全体でいかなる汎化を行ったかが不透明である。MCCSではそうした場合、「猫」、「犬」そして「亀」を同一の同値類 { 猫、犬、亀 } に分類し、それらの共通極小汎化である「脊椎動物」を選択的に選ぶ操作を行う。

列全体で同一視すべき語を決定する語の汎化選択関数 f は複数個存在し、それぞれが異なる「汎化コスト」を有する。すなわち、θ から決まる同値類の集合 $\{[w_1], ..., [w_k]\}$ に対し、$gcost(\theta, f)$ を $f([w_j])$ と $[w_j]$ の各要素をEDR辞書において接続する最小のパス長の最大値として定める。これは、同値類中の語 $w \in [w_j]$ を $f([w_j])$ で置き換えるための汎化コストを意味する。同値類は複数個存在するので、これらの最大値で $gcost(\theta, f)$、すなわち、汎化選択関数 f を用いたときの op-選択全体で要するコストと定める。$gcost(\theta, f)$ が高い場合、汎化語の抽象度は上がり、適切な汎化をなしたとはみなしえない。この立場から、本研究では汎化コストの上限値 gl を設け、gl 以内のコストを持つ語の汎化選択が可能な op-選択のみを考える。

[Definition 3]　**（極大類比）** op-選択 θ の汎化コストを $gcost(\theta) = \min\{gcost(\theta, f) | \text{mst 関数 } f\}$ で定める。所与の汎化コスト上限値 gl に対し、$gcost(\theta) \leq gl$ なる θ の中で、イベント対の集合として包含関係に関して極大なものを、極大 op-選択と呼ぶ。特に、さらに、極大 op-選択 θ とコスト条件 $gcost(\theta, h) \leq gl$ を満たす汎化選択関数 h に対し、$MCCS(\theta, h)$ を極大類比として定める。 □

定義から、極大 op-選択にイベント対を追加してできるどの op-選択もコスト条件を満たさない。また、ひとつの極大 op-選択 θ に対して、$gcost(\theta, h) \leq gl$ なる mst 関数 h は一般に複数個存在することに注意する。

[Proposition 1]　**汎化コストの単調性**　op-選択 θ_j に対し、

$\theta_1 \subseteqq \theta_2$ ならば $gcost(\theta_1) \leqq gcost(\theta_2)$ が成立する。

4. 極大類比のボトムアップ探索法

本節では、命題 1 に基づき、最小指示度 minsup のもとで頻出アイテム集合を求める APRIORI [1] とよく似た枝刈り規則を持つ、極大 op-選択のボトムアップ探索法を与える。基本的には生成ーテスト法であり、op-選択をイベント対の要素数の順で生成しながら、汎化コスト条件のテストを施す。このために、重複のない op-選択の生成順序を以下で定めておく。

[Definition 4] 文書 $D_\ell = \{g_i^{(\ell)} | i = 1, ..., n_\ell\}$ 中のイベント間に、D_j における出現順序 $\prec$ を与え、$i < j$ に対し、$g_i^{(\ell)} \prec g_j^{(\ell)}$ を仮定しておく。さらに、イベント対の集合である op-選択は、第 1 成分のイベントの順で整列した列とみなす。すなわち、

$$\theta = < g_{i_1}^{(1)}, g_{j_1}^{(2)} >, ..., < g_{i_k}^{(1)}, g_{j_k}^{(2)} >, \text{ただし、} g_{i_m}^{(1)} \prec g_{i_{m+1}}^{(1)}$$

この準備のもとに、op-選択間の direct successor 関係 $\prec$ を以下で定め、その推移的閉包を改めて $\prec$ と記す。

$\theta_1 \prec \theta_2$ iff $\theta_1 = \theta P_{ij}$, $\theta_2 = \theta P_{ij} P_{xy}$
ただし、$i < x$, $j < y$, かつ θ はある op-選択.

□

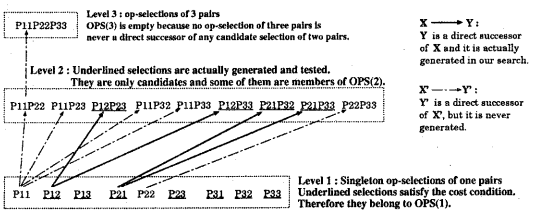

図 5 op-選択の順序構造と枝刈り（3イベントからなる文書対の場合）

定義から、所与の op-選択 θ_2 の直前の op-選択は、θ_2 の prefix として唯一に定まることに注意する（図 5 を参照せよ）。つまり、すべての op-選択からなる空間の中で、生成パスは唯一である。さらに、生成された op-選択の汎化コスト条件をテストし、テストに失敗した op-選択のどの successor も自動的にコスト条件を満たさず、したがって、枝刈りできる。下記で示す op-選択の帰納的構成は、こうした生成とそれに付随する枝刈りを用いている。

単一イベント対の構成：文書 D_1 中の i-th イベントと D_2 中の j-th イベントの対 P_{ij} でコスト条件を満たすものだけをリストアップする。

$$OPS(1) = \{P_{ij} | gcost(P_{ij}) \leqq gl\}$$

イベント対の追加処理： k 個のイベント対からなり汎化コスト条件を満たす op-選択から、単一イベント対を追加・テストを行い、次のレベルのコスト条件を満たす op-選択を求める：

$$OPS(k+1) = \{\theta P_{ij} P_{xy} \mid \theta P_{ij} \in OPS(k), P_{xy} \in OPS(1), i < x,\ j < y,\ gcost(\theta P_{ij} P_{xy}) \leqq gl \}.$$

停止条件： $OPS(k)$ の構成を $OPS(\ell) = \phi$ となるまで繰り返す。ℓ は、D_j 中のイベント数を n_j として、高々 $\min\{n_1, n_2\}$。

5. 非類似文書とのネガティブマッチング

前節までで述べた技法は、2個の類似文書の場合である。3個以上の文書に対しては2個の場合の op-選択構築技法を繰り返すことにより実現する。算出された op-選択およびその極大類比のサイズは、元の文書のサイズを超えることは決してない。したがって、繰り返しのステージが増えるにつれ、複雑さは減少する。このように、本構想の成否は、最初の文書対に対してどれだけ効率的に汎化コスト条件を満たす op-選択を自動構成できるかに依存している。そこで、非類似文書群の提示のもとで、非類似文書をカバーしない極大類比のみを生成する手法を本節で与えておく。すなわち、非類似文書をカバーする極大類比を検出し、棄却する方式である。

[Definition 5] **（非類似文書とのネガティブマッチング）** 所与の非類似文書群 $\mathcal{D}_{non} = \{N_1, ..., N_m\}$ と類似文書群から抽出された極大類比 $G = g_1, ..., g_k$ を考える。このとき、G は下記の条件を満たすとき N_j を包摂するという。

N_j 中のイベント列 $s_{i_1}, ..., s_{i_k}$ で、$s_{i_j} \preceq g_j$ $(j = 1, ..., k)$ なるものが存在する。

G は、$\mathcal{D}_{non}$ 中のどの文書も包摂しないとき、ネガティブマッチングに成功するという。 □

$G = g_1, ..., g_k$ が $N_j = s_1, ..., s_m$ を包摂するか否かのテストは下記の最小化演算子 μ_j を用いた簡単な方法で実行できる。

$$c_1 := \mu_j\ (s_j \leqq p_1),\quad c_{n+1} := \mu_j (s_j \leqq p_{n+1} \text{ and } j \geqq c_n + 1)$$

[Proposition 2] $G = \{g_1, ..., g_k\}$ が非関連文書 N_j を包摂する iff $c_1, ..., c_k$ が構成可能（包摂しないときは、ある j に対し、μ_j が存在しない）。

命題および、イベントの包摂関係のテストが木のサイズに関する多項式時間で実行可能なことから、ネガティブマッチングは多項式時間で計算可能である。この単純な事実から、次節においては、非関連文書群とのネガティブマッチングは行わず、2文書からの極大類比形成過程の計算量と品質の評価のみに注目した実験結果のみを示しておく。

ネガティブマッチングに関連した別のアプローチとして、どの非類似文書もカバーしない概念木の列を構成する問題を考えることも可能である。実際、石川の研究 [4] では、概念木ではなくより一般の概念グラフに対して、その正例負例のもとでの帰納的汎化の問題を解いている。最悪の場合の計算量は指数オーダーだが、対象をより単純な概念木に限定できるので、より効率的なアルゴリズムを設計する可能性は残っている。

6. 実験結果と展望

３０個程度の文から構成される２つの童話（図 6）に対して、イベント対数が最大となる極大 op-選択を全て求める実験を４Gバイトメモリ上で行った。

２つのお話はともに、

(P1) ２人の兄弟がおり、(P2) 弟はある貴重なものを

むかしむかし、二人の兄弟が山に住んでいました。兄弟は江戸に働きに行きました。弟は熱心に働きました。...	王は猪を倒した者に娘をあげると約束しました。王国に兄弟がいました。兄弟は猪を探しました。...
『歌う骸骨』	『歌う骨』

図 6　実験に使った物語

手にいれるが、(P3) 兄によって殺されてしまう。(P4) 骸骨となった弟は兄の悪事を暴く歌をうたい、その結果、(P5) 兄は罰をうける。

という共通性をもっている。現在のテキストからイベント（の概念グラフ表現）を作成する前処理においては、照応解析、深層格の抽出といった意味処理を行っておらず、したがって、上記の共通性のうち、(P1), (P3) および (P4) のみが検出された極大類比によってカバーされるにすぎない。(P2) に関しては、『歌う骸骨』においては貴重品とは「記念品」であり、『歌う骨』では「猪」であり、ＥＤＲにおいて共通性を見出すには離れすぎた単語である。実際、汎化コストの上限値は４で実験してみたが、「猪」と「記念品」を同一視する同値類は、決して形成されなかった。

この結果から、通常の辞書的意味での近さ・遠さでなく、役割の近さ・遠さが必要となることが示唆される。つまり、カテゴリーは離れていても、似た位置づけが可能である場合は、汎化コスト条件を緩和させる必要性があるだろう。もちろん、全く異なる概念が同様な役割を持つか否かの判定は、高度な解釈問題である。現在のところは、そうした解釈の可能性を含むものを汎化コスト条件で排除しないという意味での条件緩和を考えている。すなわち、重要な項と同じ深層格で結合されたものに対してより緩やかな汎化コストを認める戦略である。このために、項の重要度を付与する操作が必要になるが、これに関しては、文書要約やキーグラフの技法 [6], [7] が使えると期待している。

次に (P5) に相当する文は、『歌う骸骨』では『兄は刑務所に連れていかれました』、『歌う骨』では『兄は水の中に投げ込まれました』である。後者では文脈において「罰」を表している。また、「連れていく」も目的語が「刑務所」で初めて「罰」の意味あいを持つ。現在のところ、こうした解析は行っておらず、意味処理・文脈処理の深みにどこまで踏み込むべきかを、対象とする文書の種類も含めて思案中である。

最後に、計算量に関する考察を述べておく。展開された全 op-選択数は、５万個程度である（図７）。また、汎化コストの他に、極端に長さが異なる文対の対応を防ぐ追加的なパラメータを用いているが、その影響はそれほど大きなものではない。汎化コストの上限値は４である。これは、４以上の上限値では、抽象度が極めて大きな動詞への汎化が行われ、意味をなさないイベント対が形成されることによる。ただし、動詞と名詞では、辞書の粒度が全く異なり、品詞毎に汎化上限を設けることを予定している。５０個程度のイベント数になると、現在の試験的プログラムではオーバーフローしたが、同一文書中の文間の類似関係を利用することにより、代表的な極大 op-選択の生成には成功した。ここで、ひとつの文書中の類似したイベントとは、そのＭＣＳが低い汎化コストで構成できることを意味しており、op-選択の生成過程において、既に対応付けられたイベントと類似したイベントに対する対を op-選択に追加しない処理を行っている。

上記の実験結果と考察に基づいて、近い将来に、必要最低限度の深層格の処理と、重要度に基づく汎化コスト制約の緩和により、事実の記述を主とする１００文程度の文章群に対する極大類比構成手法の実現を目指している。文脈に応じた表現の違いの問題に対しては、明確な計画を現時点ではもっているわけではないが、語彙の類似性のみならず表現の類似性を計測する研究もあるので、そうした技法をできる限り取り込むことにより、この困難な問題にアタックしたいと考えている。

イベント対数	実際に展開された op-選択数	理論上可能な全 op-選択数
1	39	900
2	868	189,225
3	5712	16,483,600
4	14916	751,034,025
5	17285	20,307,960,036
6	9246	352,568,750,625
7	2254	4,144,481,640,000
8	240	34,256,731,055,625
9	0（停止条件）	....

図７　３０イベントからなる２文書の処理

謝　辞

本研究は未来開拓研究「情報知財の組織化とアクセスの感性的インタフェース」のサブテーマである「物語データベース」の実現を目指して実施している。代表者の北大・田中譲教授、サブテーマ責任者の北大・山本章博助教授に感謝したい。

文　献

[1] R. Agrawal, R. Srikant: Fast Algorithms for Mining Association Rules, Proc. of the 20th Int'l Conf. on Very Large Data Bases, 478–499, 1994.

[2] Cohen,W.W. & Hirsh,H.: The Learnability of Description Logics with Equality Constraints, *Machine Learning*, Vol. 17, No. 2–3, pp 169–199 (1996).

[3] ELECTRONIC DICTIONARY VERSION 2.0 TECHNICAL GUIDE, TR2-007, Japan Electronic Dictionary Research Institute, Ltd. (EDR), 1998.
http://www.iijnet.or.jp/edr/

[4] 石川裕治: 演繹オブジェクト指向データベースにおける知識修正のための帰納的手法の研究, 平成７年度修士論文、東京工業大学総合理工学研究科システム科学専攻 (1996).

[5] McKeouwn99, K.R. et.al.: Towards Multidocuments Summarization by Reformulation: Progress and Prospects, Proc. AAAI99, 453–460 (1999).

[6] Y. Ohsawa, N. E. Benson and M. Yachida: KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor, *Proc. of IEEE International Forum on Research and Technology: Advances in Digital Libraries - ADL'98,* pp. 12–18, 1998.

[7] K. Zechner: Fast generation of abstracts from general domain text corpora by extracting relevant sentences, *Proc. of the 16th Int'l Conf. on Computational Linguistics,* pp. 986–989 , 1996.