社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－NL－162 (2)
2004／7／15

# クエリー駆動方式の文書間極大類比構築法

原口 誠　　吉岡 真治

北海道大学情報科学研究科コンピュータサイエンス専攻

**概要**：様々な観点から類似していると判断できる文書群から、文書群に内在する共通性を汎化された共通イベント列（極大類比）として抽出し、極大類比に包摂される任意の文書を関連文書と判断するアクセス手法を考える。極大類比を構成する際に必要な計算量を軽減せしめ、かつ同時に関連性判断の精度を向上させる目的で、所与の質問式Qの具体化で事例文書Iの汎化になっている極大類比に限定した構成法を与える。

# A Query-Driven Construction of Maximal Analogies from Documents

Makoto HARAGUCHI and Masaharu YOSHIOKA

Division of Computer Science, Hokkaido University

**Abstract:** We present here a method for constructing a common abstract event sequence from two or more documents also regarded as event sequences. To reduce the computational cost and to improve the precision of relevance judgement made by the abstract event sequences, we introduce a notion of base queries and consider a retrieval process to find an instance of base query that is similar to a given family of instance documents. By the base query and its instance documents, we can design a divide and conquer strategy for the construction of abstract common event sequence from the instances.

## 1　はじめに

様々な観点から類似していると判断できる文書群から、文書群に内在する共通性を汎化された共通イベント列（極大類比）として抽出し、極大類比に包摂される任意の文書を関連文書と判断するアクセス手法を考える。極大類比を構成する際に必要な計算量を軽減せしめ、かつ同時に関連性判断の精度を向上させる目的で、所与の質問式Qの具体化で事例文書Iの汎化になっている極大類比に限定した構成法を与える。極大類比は、類似した文書群の共通したイベント列であることから、文書群が持つストーリー性、すなわち、イベントの展開の仕方を抽象化して持つことになる。したがって、こうした極大類比を拡張されたインデッキシングもしくは文書群のラベルとして保持することは、文書の整理と検索に寄与すると考えている。

関連する研究としては、汎化された５Ｗ１Ｈ情報をラベルとして用いる[Ikeda98]や、汎化された構文解析木の断片を考える[上田他02]等の研究がある。これらに対し本研究では、格構造を表す概念グラフの、集合ではなく列を考え、汎化したレベルで共通部分列を共有できる場合は、特筆すべき類似性を有し、よって、文書群のラベルとして利用可能であるとの立場にたつ。

言うまでもなく、類似文書群のラベルを高度化させるこうしたアプローチの難点は、ラベル付けに要する組み合わせ爆発の問題である。本研究のスタートを与えた文献[Haraguchi02]においては、組み合わせに関して単調なコスト関数に基づく枝刈規則を搭載したボトムアップ極大類比構成器を示したが、５０イベント程度の２文書に対し、約１０分の計算時間を要するものであった。

本報告においては、文書間に共通した抽象イベント列の使われ方まで考慮した、極大類比の算出法を

新たに与える。

適度の抽象度を持つイベント列を質問式Qとして与える。これは、探したいものを正確には表現できないときに、対象とする文書が大まかに満たすべき制約として用いる。Qによって包摂される文書群が、潜在的にヒットする可能性を持つ文書となり、この意味で、Qは検索上限を与えると理解してよい。

Qを与える人は、厳密に何が欲しいかを指定できなかったとしても、その上限指定たるQと共に、彼もしくは彼女が欲しいものの例示を与えることはできるであろう。そうした例示文書を $I_1, ..., I_k$ とする。

こうして与えられた所与の質問 $Q$ と例示文書 $I_j$ に対し、

(R1) $Q$ であって、かつ、
(R2) $I_j$ のようなもの

を探すことがここでの問題のスペックである。つまり、(R2) によって $I_j$ と（汎化されたレベルで）イベント列を共有し、しかも、(R1) によって $Q$ よりも特殊なイベント列を含む文書がヒットすべき文書であるとする。

計算論的に述べれば、$I_j$ が共有すべき汎化イベント列は $Q$ によって包摂されなければならず、このことにより、汎化イベントを構成する際のイベント対の選択に制限をおくことができ、$I_1, ..., I_k$ 間のセグメント（イベントの連続した列）対応に限定することが可能となる（節 4）。

次節において、極大類比の定義を述べ、質問 $Q$ とその例示 $I_1, ..., I_k$ の効用は、節 4 で述べることにする。なお、簡単のために、$k = 2$ として述べるが、一般の $k$ に対しても有効である。

## 2 極大類比

物語とは因果関係などのイベント間の依存関係が記述されたものとして理解できる。例えば、『... するために ... した』などの手がかりとなる表現が明示された場合はそうした依存関係を容易に抽出できるが、どのイベントが別のイベントの前提や原因となっているかが明示されているとは限らないし、また、明示されているとしてもその書き方は様々である。このような問題を回避するために、本研究では、

**類似イベントの現れ方に関する仮定**：類似した文書には類似したイベントが同じ順序で出現する

ことを仮定する。例えば、『次郎は東京に行き、金持ちになった』という物語と『太郎は大阪に行き、富豪になった』という物語には、大都市への移動というイベントと「裕福な人」になったというイベントが同じ順序で現れており、それゆえに、『ある人が大都市にいって、裕福な人になった』という共通の汎化されたイベント列を得ることができる。後者の汎化イベント列をここでは極大類比として定める。

より正確に述べれば、まず、入力文書中の各文を形態素解析と構文解析に基づいて、語彙と係り受け・格関係からなる概念グラフに変換し、これをイベントと呼ぶ[1]。一つの文章はそうしたイベントの列として内部表現する。

次に行うべきことは、所与の２つの類似文書から、共通したイベント（部分）列を抽出することである。３個以上の類似文書からなる場合は、２個の文書に対する汎化操作を逐次的に繰り返す。ここで、汎化文書も一つのイベント列であることからそうした逐次操作が可能となることに注意したい。

２つの類似文書 $D_1$ と $D_2$ の汎化を行うために、どの $D_1$ 中のイベント $e_1$ と $D_2$ 中のイベント $e_2$ が対応する類似したイベントであるかを決める必要がある。本研究では、先に述べた類似イベントの現れ方に関する仮定にしたがって、$D_1$ と $D_2$ のイベント対の列 $< e_{11}, e_{21} >, ..., < e_{1n}, e_{2n} >$ で、その $D_j$ への射影 $e_{j1}, ..., e_{jn}$ が $D_j$ での出現順であるものだけを考え、これを**候補イベント対列**と呼ぶ。候補イベント対列の中には、そもそも類似しているとは言い難いイベント対 $< e_{1k}, e_{2k} >$ も含まれている。そうした不適切なものを排除する必要がある。このために、$e_{1k}, e_{2k}$ からの共通な汎化イベントを構成するコスト関数を概念辞書を用いて定め、ある一定のコストを超えるイベント対を含むイベント対列を排除する戦略を採用している。その際、

**コストの単調性**：候補イベント対列に新たなイベント対を追加すると、コストは

[1]現在は表層格のみの処理を行っており、当然、極大類比の品質に影響を及ぼしている。近い将来に深層格処理や照応解析も取り込む予定だが、抽出される極大類比の品質は意味解析のレベルに応じたものになる。

増加する

性質を用いて、これ以上イベント対を追加できない候補イベント対列を**極大類比**と定め、これを算出するプログラムを作成した。日本昔話「歌う骸骨」とグリム童話「歌う骨」を例題に実験を行い、２５文からなる要約文書に対して５１秒、５０文からなる要約文書では１０分程度で極大類比を算出できる。物語データベースに収録される文書が５０文だとしても、類似文書数が増加した場合を想定すると、１０分という時間は長すぎると思われる。また、そもそも１００文程度になれば、このアルゴリズムでは、候補イベント対列用のワークスペースがオーバフローしてしまう。次節ではこうした問題点を回避するための一つの試みについて述べる。

# 3 質問式とイベント列間の包摂関係

本稿で与える質問式の意味を正確に述べるために、イベント列間の包摂関係を説明しておく。前節で述べたように、各文を動詞を根に、表層格をリンクに、名詞や形容詞等の動詞以外の品詞を根以外のノードに割り当てた概念グラフ[2]で表し、各文書や質問式はそうした概念グラフの有限列として表現する。また、概念グラフをイベントの表現と考え、単にイベントという。

イベントの有限列 $AS = (e_1, ..., e_n)$ が同じく別のイベントの有限列 $AS' = (e'_1, ..., e'_m)$ を包摂する（あるいは、$AS$ は $AS'$ の汎化であるとも言う）とは、$AS$ から $AS'$ への埋め込み $\varphi$ が存在し、$e'_{\varphi(i)} \preceq e_i$ が全ての $i$ に対して成立することをさす。ここでイベント $A, B$ に対し、$A \preceq B$ は $B$ が $A$ の汎化（$A$ は $B$ の具体化）であることを記している。極大類比を、イベント列の汎化の言葉で述べなおせば、所与の文書 $I_1, I_2$ から、それらの共通汎化イベント列 $MA$ を構成する問題と言える。

**問題の定義**：イベント列 $Q, I_1, I_2$ に対し、$Q$ に包摂される $I_j$ の極大類比 $MA_1, ..., MA_\ell$ を構成し、どれかの $MA_m$ に包摂されるイベント列（文書）を質問式 $Q$ と例示文書 $I_1, I_2$ に関連した文書だと定める。

質問との関連性における代替定義としては、$MA_m$ に結束性や一貫性の観点からスコアーをつけ、スコアーの高い極大類比のインスタンスのみを関連文書と定める方式もあるが、これについては今後の課題としておく。

[2]実際には木であるが、「概念木」とよぶと別の意味も派生するので、ここでは、知識表現における標準的な用語である「概念グラフ」を用いる。

# 4 分割統治戦略

前節で与えた問題の定義により、極大類比 $MA$ は $Q = (ae_1, ..., ae_n)$ の具体化であり、ある埋め込み $\varphi$ に対し、

$$MA = (..., ma_{\varphi(1)}, ..., ma_{\varphi(2)}, .... ma_{\varphi(n)}, ...)$$

と書ける。さらに、$I_1, I_2$ は $MA$ のインスタンスであることから、各 $I_j$ は、埋め込み $\psi_i$ を用いて

$$I_j = (...., e_{\psi_i(\varphi(1))}, ..., e_{\psi_i(\varphi(2))}, ..., e_{\psi_i(\varphi(n))}, ....)$$
$$e_{\psi_i(\varphi(j))} \preceq ma_{\varphi(j)} \preceq ae_j \quad (\text{全ての } i, j)$$

となる。すなわち、

(P1) 質問式 $Q = (ae_1, ..., ae_n)$ の全てのイベント $ae_j$ に対して、各例示文書 $I_j$ は対応するイベント $e_{\psi_i(\varphi(j))}$ を必ず含み、

(P2) 極大類比 $MA$ は、$e_{\psi_i(\varphi(j))}$ と $e_{\psi_i(\varphi(j+1))}$ で挟まれたイベントセグメントの共通汎化を部分列として持つ。ただし、$e_{\psi_i(\varphi(0))}$ は $I_i$ の先頭イベントとする。同様に、$e_{\psi_i(\varphi(n+1))}$ は $I_i$ の最後のイベントとする。

上記の２つの性質 (P1), (P2) より、極大類比の構成のためには、

(C1) 質問式Qから各例示文書 $I_i$ への埋め込みを求め、$\psi_i(\varphi(j))$ で挟まれた $I_i$ のセグメント $SEG_{i,j}$ を定める。

(C2) $I_j$ の各セグメント $SEG_{i,0}, ..., SEG_{i,n}$ に対し、対応するセグメント $SEG_{1,j}$ と $SEG_{2,j}$ を所与の２文書と考え、2節で与えたアルゴリズムにより、(部分的な) 極大類比 $MA_j$ を構成し、最後にこれらを連接してできるイベント列を最終的な出力とする。

平たく言えば、質問式から例示文書への可能な埋め込みを先に計算し、イベントのセグメントを求め、セグメントに限定した極大類比形成を行う分割統治アルゴリズムである。

## 5 セグメント切り出しとイベントグラフ

前節で述べたセグメントを切り出しは、計算論的には、拡張パターンマッチングとイベントグラフなる特殊なグラフにおけるパス探索の問題に帰着させる。

まず、質問式 $Q = (ae_1, ..., ae_m)$ を、文字列パターンを概念グラフパターンに拡張した $ae_1 * ae_2 * ... * ae_m$ と捉え、各事例文書（これも概念グラフを文字とみなした「文字列」とする）に対して走査する。ただし、パターン中の $*$ はワイルドカードで、任意の概念グラフとマッチする。「概念グラフ同士の一致」は、質問式中のイベントが事例文書中のイベントの汎化になっているとき、「一致する」と考える。こうした拡張マッチングにより、所与の質問式 $Q$ に対して各事例中のイベント $e$ がマッチしうる質問式中のイベント集合 $Q(e)$ が決まる。セグメント切り出しを効率的に行うために、質問式中のイベント $ae_\ell$ 毎に、事例文書のイベント対（一般にはイベントタプル）$< E_{1,n_1}, E_{2,n_2} >$ で $ae \in Q(E_{i,n_i})$ $(i = 1, 2)$ なるものを作成する。つまり、質問式中のイベントの具体化の対の候補集合 $stratum(ae_\ell)$ を、先に形成しておく。イベント対グラフは、事例文書のイベント対をノードとし、$stratum(ae_\ell)$ で「層別」されたものとして定める。すなわち、

(C1) 各層内のイベント対はエッジをもたない

(C2) 隣接した層 $stratum(ae_i)$ と $stratum(ae_{i+1})$ において、$n_1 < m_1$ かつ $n_2 < m_2$ のとき、またそのときに限り、$< e_{1,n_1}, e_{2,n_2} > \in stratum(ae_i)$ から $< e_{1,m_1}, e_{2,m_2} > \in stratum(ae_i)$ へ（有向）エッジを張る。

ただし、(C2) の条件は、極大類比を構成するイベント対列が持つ順序制約から派生していることに注意する。

**イベント対グラフと極大類比の関係**：質問式と事例文書に対する極大類比は、上記で構成したグラフにおける初期層 $stratum(ae_1)$ と最終層 $stratum(ae_n)$ を結ぶパス上に現れるイベント対を必ず含むイベント対列から構成できる

上記のパス

$< e_{1,m_1}, e_{2,n_2} >, < e_{1,m_2}, e_{2,n_2} >, ..., < e_{1,m_x}, e_{2,n_x} >$

に対し、文書 $I_1$ で $\{e_{1,m_j+1}, e_{1,m_j+2}, ..., e_{1,m_{j+1}-1}\}$、および文書 $I_2$ に対しては $\{e_{1,n_j+1}, e_{1,n_j+2}, ..., e_{1,n_{j+1}-1}\}$ がイベントグラフから構成されるセグメントとなる。

パスの総数が増える場合は、セグメントの可能数や生成される最終的な極大類比の数が増加する。この問題点を解決するために、イベント対グラフのエッジに重みをつけ、パスの重み（エッジ重みの線形和）が小さな順に、所与の与えられた個数だけ選択的に求める方式を採用している。

## 6 実験結果と問題点

各々が４６イベントからなる２文書に対し、質問式とそれに伴う分割統治を行う場合と、分割統治を行わない [Haraguchi02] の２者を比較した。処理時間については後者が１０分であるのに対し、前者が約３秒という大幅な改善が達成され、極大類比の数に関しても、後者の約２万個に対し、前者は１１７８個に削減することができた（つまり、２万マイナス１１７８個のイベント対列が質問式の具体化ではなかったということ）。

この結果は今後に希望を見出せる結果だとも言えるが、実際問題としては、質問式の与え方が厳しすぎる場合は、そもそも所与の事例文書が質問式の具体化にならない場合もありえる。また、概念グラフおよびイベント列間の包摂関係のチェックのために電子化辞書を使っているが、電子化辞書の粒度が「ふぞろい」であることから、包摂関係の定義を緩和させ、厳密な意味では具体化になっていない場合でも具体化になっていると見做す処理も必要となる。これらの処理の組み込みは今後の課題である。

## 参考文献


[上田他 02] 上田、小山：共通意味断片の抽出による複数文書要約、言語処理学会第６回年次大会、360–363, 2000.

[Haraguchi02] M.Haraguchi, S.Nakano and M.Yoshioka : Discovery of Maximal Analogies between Stories, Springer LNAI 2534 (Proc. 5th Int'l. Conf. Discovery Science), 324-331, 2002.

[Ikeda98] T.Ikeda, A.Okumura, K.Muraki: Information Classification and Navigation Based on 5W1H of the Target Information. COLING-ACL 1998. 571-577