構造類比 － 解釈の多様性と潜在的共通構造の発見に向けて

原口誠

北海道大学大学院情報科学研究科コンピュータサイエンス専攻

Abstract: ＡＩ研究が始まって以来、類似性に基づく推論としての類推は、日常的推論の世界において類推が創造性に直結しているとの直感もあり、様々な角度から研究がなされてきた。類似性の検出問題は、可能な類似性の探索空間の中での探索問題として定式化されてきたが、これは問題の定義から計算論的な困難さを内包している。したがって、人の創造的活動を支援する立場からは、検出というよりむしろ知識の「創造的整理」に寄与する理論と技法の方がより現実的との指摘もある。　本講演においては、後者の立場から、所与の文書を多様な類似性を留保しつつ階層的に組織化する試みについて述べる。キーとなるコンセプトは、イベント列の汎化による文書の構造類比である。類似性の多様性をある程度認める必要性から言うまでもなくこのアプローチも計算論的困難さを含んでいるが、データマイニング的手法、ピンポイントクラスター解析を用いた潜在情報抽出などを組み合わせることにより、文書間の重要な共通構造を抽出し、構造が持つ自然な包摂関係に基づく階層を見出す手法の現状と課題をレポートする。

1　はじめに：物語データベース構築に向けて

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、主題やトピックに基づいた文書分類に関する研究が精力的に遂行されている。膨大さと複雑さを回避するための分類とは、一般には、階層化を意味し、階層的クラスタリングに準じたものが対応すると考えられる。階層的クラスタリングでは、ある基準にもとづいたクラスの分離や統合のやり方が定められており、分類の多様性を犠牲にするかわりに、大量データにも対応できる分類を高速に実現している。しかし、得られた分類に満足できない場合は、手法の良し悪しの吟味とともに、距離や分類基準に立ち戻って検討する必要がある。例えば、「視点に基づいた動的な距離関数」があれば、異なる分類基準へのシフトとそれに基づく再分類が容易になるが、そもそも、そうした便利な距離の設計が容易でないことも明らかである。したがって、ある程度のコストを覚悟の上に、将来使うかもしれない別の観点や視点からの再分類に対応できる「構造」を予め作成しておくことも、少なくとも検討の余地があると思われる。

本研究では、この観点から、文書データベースを物語データベースとして構築することを試みている。すなわち、各物語（文書）を複数の方法で汎化し、構造的類似性に基づいて汎化された物語を拡張インデックスとして物語へのアクセスのために用いる。あたかも、複数のキーワードを文書に付与し、それらを介した文書の検索とアクセスを行うようにである。問題は、可能な汎化の数もしくは構造類比の数が膨大であり、かつ、一つの汎化を作成するタスクも本質的には組合せ爆発問題であるという事実にある。組み合わせ爆発を抑制するための枝刈り規則や文書の型や種類に特化した経験則の導入も必要となるが、これらと平行して、文書そのもののサイズを減らす操作も重要となる。つまり、重要でない文書中の文やイベントを、汎化の対象から除外し、重要な文やイベントに絞りこんだ要約文書群から、それらのイベント列汎化によって構造類比を抽出する戦略をとるとする。

さて、ここで常に問題となるのは、何をもって重要と見做すかである。素早く結論だけ知りたいのであれば、中身の物語性は無視して結論部を同定し、そこからのみ重要文を抽出するのが手っ取り早い。要約と検索の目的がこのようなものであればそれで十分だろう。一方、結論は同じでも結論の導出の仕方に興味のある人にはそうはいかない。前提から結論を結びつけるストリーに依存した要約と検索が必要となる。ストリー性を保障する重要度の定義においては、文と文の（単語）を介した結束性のみならず、話題と話題を結合する機能を持つ文もしくは語が重要な役割を果たす。シーンの切れ目や伏線となるようなものもそうした機能を有している。そこで、本稿ではそうした

物語としての展開構造の保存性

を要請し、それを保障するための要約についても述べる。

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（文間深層格として）比較的に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 [Haraguchi02, Yoshioka05]。ちなみに、これは [Haraguchi85] における極大類比を列の汎化に拡張したものとして理解できる。

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。図１における<g11,g21>,<g12,g22> がその例である。

[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。

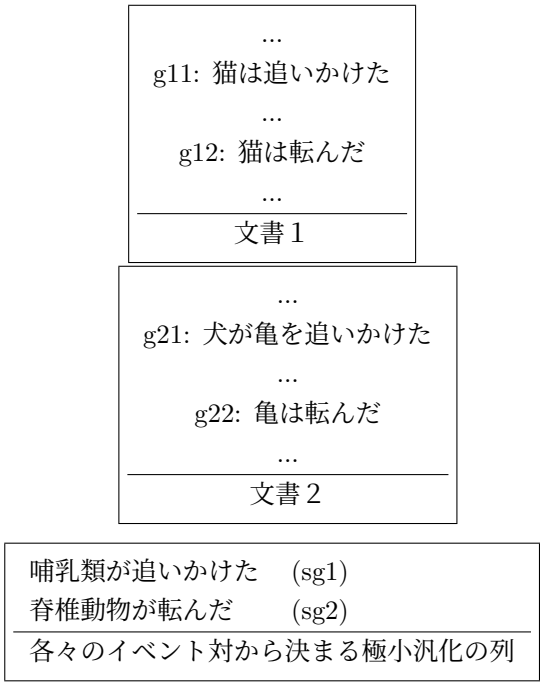

図 1: イベントとその列の汎化

イベント対列中の各イベント対は、他のイベント対とは独立に汎化も行うことも考えられる。この例の場合は、「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。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。そこで、「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（図 1. を参照のこと）。すなわち、

> **整合性条件：** イベント列において、同一の語彙が複数の異なる上位語に汎化されることを禁止する。

を要請し、イベント列の可能な汎化に意味的な制約を付与する。このようにして、異なる語彙を用いる異なる文書内のイベント対に対し、列として汎化操作を続けていけば、上位語に置き換える操作としての汎化コストは単調に増大していく。過度の汎化を避ける観点から、データマイニングにおいて標準的・古典的な、「単調性に基づく枚挙の枝刈 [Agrawal94]」をここでも行う。すなわち、

> **コストの単調性と枝刈り：** 整合性条件を満たす候補イベント対列に新たなイベント対を追加すると、コストは単調に増加し、その値が一定の値を越えたときに、汎化を打ち切る

日本昔話「歌う骸骨」とグリム童話「歌う骨」を例題に実験を行ったが、２５文からなる要約文書に対して５１秒、５０文からなる要約では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、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。したがって、より重要なイベントのみを抽出し、重要なイベントの対だけからなる極大類比を構成することが必要となる。

3 Ranking アルゴリズムの導入

物語としてのイベント列を考えた場合、イベント同士は、共起語を介してその重要度の伝播を行っていると考えるのは、それほど不自然ではないと思われる。重要度は正規化すれば確率となるので、確率スコアーを求める手法は本研究で求める手法の少なくとも雛型にはなりうる。この観点から、Web のページランキングを求める PageRank [Page98] の拡張により、共起語を介したイベントの重要度伝播を行う手法の導入を試みた。拡張の要点は、イベントを Web におけるページにみたて、リンク先への推移確率 $\frac{1}{N}$（ただし、N はリンク先のページ数）を、文間の共起語により定義できる確率（分配比）で定め、**重要度の分配過程**とみなす。さらに、あるページ p の重要度 $R(p)$ を参照元のページの重要度の総和で測る式を**イベントの重要度の集積過程**と解釈する。すなわち、第１のモデルとして下記を考える。

重要度の分配： $\alpha_{p,q}$ を p から $q(\neq p)$ への**分配比**とする。つまり、$1 = \sum_{q(\neq p)} \alpha_{p,q}$，$\alpha_{p,p} = 0$ を満たし、p から q へは $\alpha_{p,q}R(p)$ が分配される。

重要度の集積： イベント p の重要度は、p と共起語を少なくとも一つ共有する q から受け取る重要度の総和である。

$$R(p) = \sum_{q \text{ と } p \text{ は共起語を共有}} \alpha_{q,p} R(q) \quad (1)$$

実際問題としては、分配比 $\alpha_{p,q}$ の定め方で、モデルは様々な挙動を示す。最も基本的なものとしては、共起語 w の語としての重要度 $score(w)$ を用いて、

$$\alpha_{p,q} = \frac{\sum_{\text{"}w \text{ は } p \text{ と } q \text{ の共起語"}} \frac{score(w)}{w \text{ を含む文の数} - 1}}{\sum_{\text{"}w \text{ は } p \text{ 中の語で他の文と共起"}} score(w)}$$

であたえられよう。ここで、分母は他の文との情報伝達に使われる p が持つ語の重要度の総和である。また分子は、文 q は文 p と共起語 $w_1, ..., w_k$ を介して p から重要度を受け取るが、各 w_j による分配は、w_j を含む文の数に（凡そ）反比例することを表している。これは、高頻出語は当たり前すぎて、他の文にそれほど重要な情報を与えないことを考えれば、妥当な定義であろう。むろん、語を共有しなければ $\alpha_{p,q} = 0$ であり、特に、どの文とも語を共有しない文は孤立文として予め削除されていることを仮定する。

さて、このようなモデルは、文章を読解しイベント間の関連を読み解く行為が、共起語を介して行われているという直感に合致しており、ある程度の妥当性を持つと思われるが、実は下記の単純な事実が成立する。

$$R(p) = \sum_{w \text{ は } p \text{ 中の語で他の文と共起}} score(w)$$

この事実は、単純な語の重要度 $socre(w)$ だけを考慮する限り、どのような評価関数 $score$ を用いたとしても、語の重要度の線形和で重要文抽出を行う要約手法と殆ど同じ効果しか持たないことを証明している。

4 関節語・関節イベントに着目したバイアスの付与

実際の Page-Rank においても、例えばトピック依存性を考慮したランキング [Haveliwala03] を可能ならしめるために、下記に示す式でランク計算が行われている。

$$R(p) = \alpha(\sum_{\text{q と p は共起語を共有}} \alpha_{q,p} R(q)) + (1-\alpha)Q(p) \quad ... (2)$$

ここで第２項は、他の文との語の共起とは独立に、各文は固有の重要度 $Q(p)$ を持ち、定常的に重み $1-\alpha$ で重要度の供給を受けることを意味している。物語の展開構造を知る上で重要な文に適切な $Q(q)$ を付与できれば、共起語を介した結束性と展開構造上の重要性を共に加味した重要度のモデルを提供できることなる。

$Q(q)$ に対しては、様々な立場からの定義が可能であると思われるが、どのような方法・モデルを用いたとしても留意すべき下記の事実がある。

> 単純マルコフ連鎖と語の共起を用いる限り、高すぎる推移確率をもたらす不適切な高頻度語の影響を抑制する技法が必要である。しかし、だからと言って、高頻度語を無視するわけにもいかない。主要な登場人物など、高頻度語にはそれなりの重要度があるからである。

この問題を解決するために、近年、筆者の研究室で試みていることを文献 [舘林 06] にしたがって紹介しよう。物語の展開において重要となるのは、単なる重要語を多数含む文のみではなく

> 文の塊であるシーンや、そうした塊を連結する語もまた重要である。

ここで求めたいものは、時系列において近接しかつ意味的にも関連性を持つイベントからなる塊であり、主題や副題となる頻出語をある程度含むものである。基本的にはキーグラフ [大澤 99] における土台に相当するものをイメージしているが、全文書から構成される土台と、特定のシーンにおける土台を比較すると、後者の方がより稠密なグラフとなり、クリークもしくは擬似クリークを形成する可能性が高まる。この理由により、

> **セグメンテーションとクリーク探索の併用手法による塊（土台）検出：**物語におけるシーンの切れ目を算出するために、テキストタイリング [Hearst94] によってセグメントをまず算出する。次に、各セグメント内において、時間的な近接関係と語の共起性に基づいてイベント間にリンクを張った非有向グラフを構成し、イベントの語の頻出性に基づくスコアーの総和が高いクリークを求める。

重み最大クリーク [Tomita03]、もしくは、重みスコアの降順にトップN個の良い塊を高速に算出するアルゴリズム [Okubo05] は良く知られており、こうしたものを用いれば上記のタスクはきわめて高速に処理できる。

次に、そうしたイベントの塊により共有される語は、例え頻度が少なくとも重要なものと認識できる。これはキーグラフにおける「屋根」の考え方そのものである：

> **関節語：** 上記のクリークをキーグラフにおける土台にみたて、関節語を抽出し、そうした関節語の含まれ具合に応じてイベント p を定常的に訪れる確率 $Q(p)$ を定める

最後に、α のチューニング手法を決めておく必要がある。現在は単純に、訓練事例に対して正答率を最大化する α を用いている。

上記を実装し、新聞の論説記事に対する実験を行った。結果を要約すると、要約を評価するときの定番である「圧縮率」により評価すると、圧縮率が厳しい場合（実験では１５%)、物語の展開構造を知る上で必要な関節語やそれを含む関節イベントが要約に掲載される率は極めて小さくなる。これは定義から当然である。一方、３０%にすると、関節語や関節イベントの識別率は向上し、殆ど全ての文書に対して正答率は向上した。

5 おわりに

前節で述べた結果に基づいて、本稿においては、下記の事実を指摘しておきたい。

1. 物語データベース構築のために必要な要約を考察する、もしくは評価する再には、圧縮率以外の、構造評価のための定性的な枠組みが必要である。
2. 意味的な塊とそれらを繋ぐ関節イベント・語を捉えるための他の手法の可能性、例えば、セグメント内クラスタリングとクリーク探索手法を質的に比較し、関節を見出すためのより的確な手法に進化させる必要がある。

参考文献

[大澤 99] 大澤 幸生・N. E. Benson・谷内田 正彦：KeyGraph: 単語共起グラフの分割・統合によるキーワード抽出，電子情報通信学会誌論文誌, J82-D1, No. 2, pp. 391 - 400, 1999.

[舘林 06] 舘林俊平、意味的構造と文間の相互依存関係に基づく文書要約手法の提案、平成 17 年度修士論文、北海道大学大学院情報科学研究科コンピュータサイエンス専攻 (2006).

[Agrawal94] 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.

[Haveliwala03] Taher H. Haveliwala: Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, pp. 784-96, 2003.

[Haraguchi85] Haraguchi,M.: Towards a Mathematical Theory of Analogy, *Bull. Inform. Cybernetics*, Vol.21, 29–56 (1985).

[Haraguchi02] M. Haraguchi, S. Nakano and M. Yoshioka: Discovery of Maximal Analogies between Stories, Springer-LNAI 2534, pp. 324 - 331, 2002.

[Hearst94] M. A. Hearst: Multi-Paragraph Segmentation of Expository Text, Proc. of the 32nd Meeting of the Association for Computational Linguistics, pp. 9 - 16, 1994.

[Okubo05] Y.Okubo and M.Haraguchi: Finding Significant Web Pages with Lower Ranks by Pseudo-Clique Search, Springer-LNAI 3735, pp. 346 - 353, 2005.

[Page98] L.Page, et.al. The PageRank Citation Ranking: Bringing Order to the Web. The PageRank Citation Ranking, Stanford Degital Library Technology Project `http://dbpubs.stanford.edu/pub/1999-66` (1998).

[Tomita03] E. Tomita and T. Seki, An Efficient Branch-and-Bound Algorithm for Finding a Maximum Clique, DMTCS'03, Springer-LNCS 2731, pp. 278 - 289, 2003.

[Yoshioka05] M. Yoshioka, M. Haraguchi and A. Mizoe: Towards Constructing Story Databases Using Maximal Analogies between Stories, Springer-LNAI 3359, pp. 243 - 255, 2005.