3C1-03

物語性を考慮した要約に基づく類似文書の汎化

Abstracting Similar Documents based on a Text Summarization Technique

原口 誠 Makoto HARAGUCHI　吉岡 真治 Masaharu YOSHIOKA　溝江 彰人 Akihito MIZOE　四ッ谷 雅輝 *1 Masaki YOTSUTANI

北海道大学 大学院工学研究科 電子情報工学専攻
Division of Electronics and Information Engineering, Hokkaido University

To classify a document set from various viewpoints, we extract a possible abstract document from given similar documents. The abstract one is a common generalization of the original documents. Although we can present an efficient algorithm to find it under some pruning rule, it is generally very hard to perform such a task of extracting abstract documents. It is therefore necessary to reduce the numbers of events appearing in documents. For this purpose, we regard one document as a network of events, present a pseudo-PageRank model for calculating the importance of events, analyze and compare it with a standard text summarization method, and finally reports the present approach to improve the pseudo-PageRank.

1. はじめに：物語データベース構築に向けて

主題やトピックに基づいた文書分類は、近年のオンラインドキュメントの急増にともなって、精力的な研究が遂行されている。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、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 [原田 02] 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の汎化のみから構成される汎化文書を抽出するために、重要文の特定を行う文書要約手法を開発することを目標とする。

連絡先: 原口 誠，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，TEL:011-706-7106(FAX 兼)，E-mail: makoto@db-ei.eng.hokudai.ac.jp

*1 現在、(株) 日立製作所。

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 $< e_{1k}, e_{2k} >$

*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。

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、最新実装版では、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、現在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。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. 擬- PageRank モデル

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を検討している。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。

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

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（例えば [望月 02]）と殆ど同じ効果しか持たないことを証明している。

4. 改良モデル、現状報告、ならびに課題

前節の擬-PageRank においては、文中の語彙はその使われ方に依存せずに、単に、スコアーと他の文との共起現象によりその重みが決定されていた。しかしながら、主格として使われているか、あるいは目的格として現れているかでは、自ずから、重要度の分配が異なるとの考えもあるだろう。文献 [四ッ谷 03] および本報告においては、その立場から下記に示す要請を満たす分配比 $\alpha_{p,q}$ の式を様々ものに変える実験を行った。具体的には

- 共起語がどの格で生じたかによる重みづけを行う。さらに、重要度を送り出す方の格に従う場合と、受ける方の格に従う場合の２通りの考え方に応じて異なる $\alpha_{p,q}$ の式が可能である。
- 重要な情報は係り受け構造において浅い部分に出現するとの立場から、深さが深いほど低い重みづけを行う。

実験結果を総括すれば、分配比の計算式の違いによる細かな違いは確かに観察できるが、擬-PageRank や単語の重要度の線形和モデルとの劇的な違いは観察することはできなかった。その理由としては下記をあげることができる。

- 文において高頻度語を持つ文は、基本的には「ハブ」として働き、それ故に、一般には高いランクを持つ。この事実は分配比（０から１の間の数）の微調整と比較して、はるかに大きな影響を持つ。

この観察から、ISF (inverted sentence frequency) と TF(term frequency) のバランスをとり、高いＴＦ値を持つ共起語による重要度の分配と集積を大胆に抑制する必要があると思われる。現在、この考え方に基づいた実験を行っている段階であり、発表時には報告したい。これに関連した今後の課題としてさらに述べれば、ISF はそもそも情報論的な解釈を行うことができるので、コーパスを用いて「驚くべき文」には初期確率として高いスコアーを付与する方式が有望だと思われる。これも、PageRank 同様の立式と固有値計算が可能なので、計算論的にも無理のない展開であろう。

擬-PageRank モデルから本質的に飛躍し、物語としての情報・重要度の伝播を捉えるための改良は、これまで述べてこな

かった点、すなわち、共起語を介した文間の「論理的なネットワーク」を考えるのではなく、イベントの文書中の出現位置に関する距離情報を組み込んだネットワークを考察することだと思われる。すなわち、近傍にあるイベント間の分配・収集は低い分配率で行い、遠くにあるイベント間には高い分配比を与えることにより、関連する距離的に近いイベント群に高い評価値が集中することを回避し、よって、文章全体からバランスよく重要なイベントを抽出する方式である。この方式も現在、実装と実験を繰り返している段階である。

参考文献

[四ッ谷 03] 四ッ谷雅輝 et.al. 共起語を介した文間の相互依存関係に基づく重要文の多段階抽出法, 言語処理学会第９回年次大会論文集, 145–148 (2003).

[望月 02] 望月源: テキスト簡易要約器 Posum マニュアル, JAIST Technical Memorandum, IS-TM-2002-002 (2002).

[原田 02] 原田実 et al.：意味グラフのマッチングによる事故問い合わせ文からの判例検索システム JCare, 自然言語処理, 9,2, 3–22 (2002).

[Page98] L.Page, et.al. The PageRank Citation Ranking, Stanford Degital Library Technology Project (1998).