文書構造解析に基づく部分文書検索

6K－3

品川 徳秀 †　北川 博之 ‡

† 筑波大学 工学研究科　‡ 筑波大学 電子・情報工学系

1 はじめに

近年、計算機環境の発展と普及に伴い、様々な情報資源が電子化され、活用されるようになってきている。中でも、電子化文書は重要な情報資源であり、それゆえにその潜在的な数、量は膨大なものである。このため、その全体を把握する事も、個々の内容を理解する事も必ずしも簡単ではなくなっている。この問題に対し、分類や検索などのプロセスを電子メディアの特性を生かして自動化するための研究がされている。

従来の文書検索においては、個々の論文や書籍などの文献全体を一つの検索の単位として問合せ処理や処理結果の呈示に用いるシステムが多かった。しかし、近年の全文テキストデータベースの増加や電子文書データの増加に伴い、具体的に文献データのどの部分が問合せに合致するかまでを特定する部分文書検索の必要性が認識されている [1] - [5]。

部分文書検索としては、あらかじめ明示的に与えられた章や節、段落といった構造に基づいた部分文書を扱うもの [3] や、文書をその記述内容に基づいて自動抽出した話題の列として扱うもの [2], [4] などがある。しかし、いずれにおいても、問合せ処理を行なう時点では部分文書の集合は固定的であり、検索候補の選択の自由度が低くなるなど、問題点もある [1]。

本研究では、文書データ中における話題に基づいて、階層的かつ詳細に文書構造を抽出し、それを利用した部分文書検索手法を提案する。これにより、話題のまとまりの粒度をより自由に調整する事が可能になる。

以下では、まず、本手法で用いた部分文書間の類似度と構造抽出の手順を述べ、抽出した構造を利用した部分文書検索について説明する。また、転置ファイルなどの事前情報を用いた問合せ処理方法と構造抽出の省略条件について説明する。最後にまとめと今後の課題を示す。

2 文書構造の抽出

本手法における構造抽出のプロセスはボトムアップに行なわれる。対象文書の初期構造として、十分に小さく機械的に境界を設定する。これによって作られる部分文書を基底ブロックと呼ぶ。また、基底ブロックからそれを連結したものとして段階的に構成される部分文書を (複合) ブロックと呼ぶ。基底ブロックは、最小粒度のブロックとして位置付ける事ができる。

2.1 ブロック間の類似度

文書間の類似度として、次式で与えられる tf*idf によって表現された文書の特徴ベクトル間のコサイン測度がよく知られている。N 個の文書からなる文書集合 R において、出現語を j、各文書 d_i における j の出現頻度を t_{ij}、j の出現する R 中の文書数を n_j とする。

$$v_i = (v_{ij})_{j=1,\ldots,k}$$

$$v_{ij} = t_{ij} \cdot log(N/n_j)$$

$$sim(d_1, d_2) = cos(v_1, v_2) = \frac{v_1 \circ v_2}{\|v_1\| \cdot \|v_2\|}$$

本手法では基底ブロックの集合を R として扱い、上式を次のように変形したものを用いる。

基底ブロックなどの比較的小さなブロックでは内容の記述量が少ないため、その中に出現する語のみでその特徴ベクトルが十分に表現できない事がある。しかし、このような状況ではその周辺と合わせて一つの話題を扱っている可能性が高い。即ち、その周辺のブロックと補完しあう関係になる。また、記述位置が離れる程、その依存関係は薄くなる。このような特徴から、その前後に位置する基底ブロックの特徴ベクトルによって次式のように補正を行なう。

$$v_i' = v_i + w_i^{(r)}$$

$$w_i^{(r)} = \sum_{p=1}^{r} \frac{w_{F-p} + w_{L+p}}{(1+p)^{(L-F+1)}}$$

ここで、F, L はそれぞれ、対象ブロック d_i の最初と最後の基底ブロック番号とし、v_i は d_i の、w_j は基底ブロック b_j の tf*idf による特徴ベクトルとする。

2.2 文書構造の抽出手順

この手順を以下に示す。

1. 対象とする文書に、語数や文数などを基準にして十分に小さく機械的に境界を設定する。これによって得られた部分文書の列を初期ブロック列とする。
2. 次のように段階的に大きなブロックを構成する。
   a. 既知の極大なブロックの列から、最も高い類似度を示す連続したブロックの組を選択。
   b. 選択された二つのブロックを、それらを連結したブロックで置き換える。
3. 全ての基底ブロックが一つのブロックとして連結されるまで、ステップ 2. を繰り返す。

これによって、最終的にブロックの連結関係を表す二分木が得られる。この連結は局所的に類似した部分に対して優先的に行なわれるため、話題のまとまりを反映した構造となる。即ち、話題の抽象化の度合が高いほど上位の節に対応する。

3 部分文書検索

問合せに対して、抽出した構造中に現れる全てのブロックを部分文書検索の候補集合とする。この構造は話題のまとまりを反映しているため、有意な内容の部分文書が検索候補になると期待される。また、検索候補の爆発的な増加を押さえる事ができる [1]。

Passage Retrieval Based on Document Structure Analysis
Norihide Shinagawa †, Hiroyuki Kitagawa ‡
† Doctoral Program in Eng., Univ. of Tsukuba,
‡ Institute of info. Sci. and Elec., Univ. of Tsukuba

4 事前情報を利用した問合せ処理

ここでは、事前情報を利用して、構造抽出を行なう範囲を問合せに応じて必要な部分に限定し、より効率的に処理を行なうための実装法について説明する。また、中間の構造抽出を省略できる条件について言及する。

4.1 利用する事前情報

事前情報として、次の二つが利用できるものとする。

転置ファイル　従来用いられてきたものと同様のものである。これを用いる事で、語 j を含む基底ブロック b_i と、それにおける出現頻度 t_{ij} を得る事ができる。

部分構造情報　事前に行なった構造抽出の結果を幾つかの部分木に分割し、その境界の位置を保存したものである。この情報を用いる事で、結合候補の範囲を限定でき、下位の構造抽出を各範囲内に局所化する事ができる。

4.2 問合せ処理の手順

問合せ処理を行なう際にこれらの事前情報を利用して、検索候補となるブロックを絞り込んだ構造抽出の手順を以下に示す。これによって、明らかに不要な部分 (問合せとの類似度が 0 になる部分) の構造抽出を省く事ができ、これは即ち、それらに含まれる不要なブロックを検索候補からはずす事を意味する。

1. 事前情報を利用して、問合せに含まれる語を含む範囲を特定する。これらを候補範囲とする。
2. 各範囲について、それに含まれる検索候補となるブロックを構成する。
   A. 候補範囲は、その内部に局所化して下位構造の抽出を行なう。
      a. より詳細な範囲情報が利用できる場合には、ステップ 2. を再帰的に適用する。
      b. 基底ブロックのみの場合は、その範囲内に限定して通常の構造抽出を行なう。
   B. 非候補範囲については内部の構造抽出は行なわず、全体を単純に連結する。これは、どの部分ブロックも候補とはなり得ないためである。
3. 全ての範囲について構造抽出が完了したら、以降はそれらの上位の構造を抽出する。

4.3 構造抽出の省略条件

更に、幾つかの条件の下では構造抽出の省略が可能である。以下、これについて説明する。

ある範囲 R_i において、それを構成するブロック列 $(b_{ij})_{j\in K}$ $(K=\{1,\ldots,k\})$ について考える。また、検索時の問合せとして、結果を絞り込むための次のような付加条件が与えられているものとする。

- 検索件数の上限 n
- 問合せとの類似度の下限 α

この時、特徴ベクトル及びコサイン測度の定義により、

$$cos(a,q)=0 \Longrightarrow cos(a+b,q) \geq cos(b,q)$$

が成立する。この事から、次が導出される。

性質 1

$$\exists! j_0 \in K \wedge cos(b_{ij_0},q)\neq 0 \Longrightarrow \forall J \subset K \left(cos(\textstyle\sum_{j\in J} b_{ij},q) \leq cos(b_{ij_0},q) \right)$$

これにより、問い合わせ語を含む唯一のブロックである b_{j_0} が付加条件によって棄却されている場合、それを含むより大きなブロックで検索結果に採用されるものは、R_i 中には存在しない事が保証される。即ち、これ以上の R_i 内部の構造の抽出は省略可能である。

性質 2

$$\forall j \in K-\{1,k\}\,(cos(b_{ij},q)=0) \Longrightarrow \forall J_1 \ni 1, J_2 \ni k\,(J_1 \oplus J_2 = K \Longrightarrow cos(\textstyle\sum_{j\in J_1} b_{ij},q) \leq cos(b_{i1},q) \wedge cos(\textstyle\sum_{j\in J_2} b_{ij},q) \leq cos(b_{ik},q))$$

これにより、両端のブロック b_1, b_k のいずれも付加条件によって棄却されている場合、それを含むより大きなブロックで検索結果に採用される可能性のあるものは、R_i 中には R_i 全体を連結したブロックのみである事が保証される。即ち、R_i の中間構造の抽出を省き、R_i 全体を連結したもののみを考えれば良い事が分かる。

5 まとめと今後の課題

情報の肥大化により、部分文書検索は今後重要な問題となると考えられる。その際に、章や節などのあらかじめ明示的に与えられた構造を用いたり、文書を単純に話題の列として扱ったりするだけではなく、問合せ処理および結果出力の粒度をより柔軟に調整できる事が望ましい。

本研究ではこれを実現する一つの手法として、記述内容に基づいた構造抽出を行ない、それを候補集合とする部分文書検索を提案した。また、事前情報を用いる事によって、与えられた問合せに応じて構造抽出を局所化し、効率的に処理を行なうための手法について説明した。

今後は、本手法による実装の詳細な評価や、問合せの視点を導入したより柔軟な構造抽出法についても研究を行なう予定である。また、ブラウジングとの有機的な統合についても検討する必要があると考える。

参考文献

[1] 品川徳秀, 北川博之, 内容解析に基づく文書構造の自動抽出, 第 116 回情報処理学会データベースシステム研究会, 1998 年 7 月.

[2] M.A.Hearst and C.Plaunt, "Subtopic Structuring for Full-Length Document Access", Proc. of SIGIR '93, ACM, pp.59-68, Pittsburg, 1993.

[3] G.Salton, J.Allan and C.Buckley, "Approaches to Passage Retrieval in Full Text Information Systems", Proc. of SIGIR '93, ACM, pp.49-58, Pittsburg, 1993.

[4] G.Salton, A.Singhal, M.Mitra and C.Buckley, "Automatic Text Decomposition Using Text Segments and Text Themes", Proc. of Hypertext '93, ACM, pp.53-65, Washington DC, 1993.

[5] M.Kaszkiek and J.Zobel, "Passage Ritrirval Revisited", Proc. of SIGIR '97, ACM, pp.178-185, Philadelphia, US, 1997.