文書部分構造の主題間の関連に基づく文献検索*

5Y－3

大石 貴治　吉川 正俊†

奈良先端科学技術大学院大学 情報科学研究科‡

1. はじめに

文献内の内容の構造に基づく検索検索を提案する．たとえば文献 A において DB に関する話題は一部でしか述べられておらず,OS の話題が文献全体で述べられているとする．一方文献 B においても DB は一部でしか述べられていないが, 文献全体ではネットワークの話題が述べられているとする．この場合にどちらか一方だけを検索したいという要求がある．これに対しこれまでの方法は文献内の構造を考慮していない．

上記の要求に答えるため，本研究では一つの文献をさらに細かいセグメントに分けることによりそのセグメントのベクトルを作り，文献の内部構造を考慮にいれた問合せを行う方法を提案する．

2. 部分構造の主題の相互関係を考慮に入れた文献検索

一つの文献をいくつかに分解して情報を得ることで, 文献全体どうしで比較を行うよりも, 部分構造の主題の相互関係を考慮にいれた検索の方が精度の高い検索ができるのではないかと考える．

そのため, 一つの文献から一つのベクトルを作るだけでなく，一つの文献をいくつかのセグメントに分け, そのセグメントからベクトルを作る．セグメントは文献名とリージョン (連続した部分文字列) 集合の対で表現する．

検索を行う際には LSI(Latent Semantic Indexing) を使用し, ベクトル検索における効率をはかる．

2.1 Latent Semantic Indexing

LSI[4)3)2)] は Salton の SMART システムで用いられた伝統的なベクトル空間技術を上回る性能を発揮する, ベクトル空間情報検索手法である．

完全な文献集合から単語-文献行列が生成される．この行列の各成分はどの文献にどの単語が出現しているかを表すものである．この行列の特異値分解 (Singular Value Decomposition(SVD)) が計算され, 小さい値は削除される．結果の単値ベクトルと単値行列は文献と問合せの単語頻度ベクトルを, 単語-文献行列からの意味関係が保存される一方で, 使用される単語の使用変化が少なく抑えられているような部分空間に射影する．文献と問合せのベクトルの正規化された内積で余弦類似尺度を計算する．そして文献はこの計算された値である関連度 (類似度) の順に並べられる．

2.2 話題の相互関係に基づくセグメントの構築

一つのセグメントは複数のリージョンから成ってもよいが, 簡単のために以下では一つのリージョンを持つ場合について述べる．

セグメントの相関関係は対応するリージョンの包含関係や並列関係に基づく．たとえばリージョン $R_a(s_a, e_a)$,(s_a はリージョンの開始位置，e_aはリージョンの終了位置とする) の topic を A, リージョン $R_b(s_b, e_b)$ の topic を B とし, $s_a \leq s_b$かつ $e_b \leq e_a$ であるという情報を持つセグメントをそれぞれ作成すれば, 文献の中の大きな話題が topic A, その中の部分的な話題が topic B であるような, 二階層の包含関係で文献を考えることができる．

さらに, リージョン $R_c(s_c, e_c)$ の topic が C であり, $s_a \leq s_b \leq s_c$かつ $e_c \leq e_b \leq e_a$ であるという情報を持つセグメントをそれぞれ作成すれば, topic B の中の部分的な話題 topic C も考慮した三階層の包含関係で文献を考えることができる．

また, リージョン $R_d(s_d, e_d), R_e(s_e, e_e)$ の topic がそれぞれ D,E であり, $s_d < s_e$かつ $e_d < e_e$ であるという情報を持つセグメントをそれぞれ作成すれば, 包含関係でなく並列な関係で文献を考えることができる．

このようにそれぞれの文献を話題の相互関係に基づいてセグメントを構築し文献を捉えることによって, その関係を用いた検索が可能となる．

2.3 セグメントを作る方法

一つの文献をいくつかのセグメントに分けなければならないが、ここで三種類の方法を述べる。はじめの一つはヒューリスティックに構成する方法であり, 残りの二つは自動的に構成する方法である．

- セグメントの作成者が文献, 文章の内容を理解してそれに基づいてセグメントを作成する．この場合セグメントは連続していなくてもよいし, セグメント同士がオーバーラップしていてもよい．
- あらかじめ著者が章，段落などに文献を分けておく．その章，段落を一つのセグメントにする．
- まず一つの文献 (n 文で構成) に対して TextTiling[1)] でブロックの大きさを 1 から n までのすべてのパターンを行い，n 個の関連度曲線を作る．(ブロックの大きさが 2 というのは一つのブロックが 2 文からできていること示す．) 次に山と谷の関連度の差の大きさに閾値をもたせ n 個の関連度曲線からいくつかを pick up する．どのくらいの閾値をもたせるか，どのくらいの数の関連度曲線を選ぶのかは課題である．
  最後に取り出した関連度曲線から山の部分をみつけ，それを一つのセグメントとする．

2.4 問合せ

問合せの方法はそれぞれの話題の相互関係に基づいた形で行われる．たとえば二階層の包含関係の場合に

---

*Document Retrieval Based on Interrelationship of Documents' Substructures

†Takaharu OISHI and Masatoshi YOSHIKAWA

‡Graduate School of Information Science, Nara Institute of Science and Technology(NAIST)

は大きな話題に対する問合せのためのベクトルと部分的な話題に対する問合せのベクトルの二つを用いて問合せが行われる．そして，それぞれの検索の結果を組み合わせることにより答えが返ることになる．

また，三階層の包含関係であれば三つの問合せベクトルを用いて問合せが行われる．

問合せベクトルはLSIの特徴を活かして以下のように文献とキーワードを混合して構成できる．

- 文献の組み合わせから構成
- キーワードの組み合わせから構成
- 文献とキーワードの組み合わせから構成

3. 実験システムについて

奈良先端科学技術大学院大学情報科学研究科の1994,1995,1996,1997年度の修了生のうち，368人の修士論文を対象に，本論文で提案している手法を適用した実験システムを構築した．ここでは1つの修士論文全体の話題をmain topic，部分的な話題をsubtopicと読んでいる．以下にこのシステムの処理の流れについて述べる．

1. 一つの修士論文全体をmain topicのセグメントとし，それぞれの修士論文を章ごとに分けたものをsubtopicのセグメントとする．368個のmain topic用のセグメントと2389個のsubtopic用のセグメントを合わせた2757個のセグメントを集合全体とする．
2. 各セグメントに語幹の切り出し処理(stemming)を施し，不要語(stop words)を除去する[5]．
3. 文献集合で用いられている異なり単語(different term)の数を数え，文献(セグメント)-単語行列を作る．
4. 行列の特異値分解を行う．
5. 今回の実験ではmainとsubの二階層の包含関係を用いた検索を行う．まずはじめにmain topicとsubtopicそれぞれについての検索を行う．二つのベクトル$\vec{a}$と$\vec{b}$の関連度rの計算式には次式を使う．

$$r = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

main topicの関連度をm,subtopicの関連度をsとし，

$$sum = \alpha m + \beta s \quad (\alpha + \beta = 1)$$

という式を使い絞り込む．このα, βの値はユーザが指定する．さらにユーザはsumの上位何個の結果を得たいということも指定する．結果として文献の名前とその章番号と章のタイトルを返す．

3.1 実験結果

maintopicが“映像データベースのための論理データモデルとその実装”(文献)，subtopicが“時間”(キーワード),“軸”(キーワード)，“映像”(キーワード),“演算”(キーワード)，“合成”(キーワード),“場面”(キーワード),“問合せ”(キーワード)，$\alpha = 0.5, \beta = 0.5$とした場合の実験結果を表1，表2に示す．

	文献番号	章番号, 章名	関連度	適合
1	9351103	2, 映像オブジェクト	1.555	○
2	9451024	3, 映像の論理構造	1.499	○
3	9351103	4, 映像の物理的な格納構造	1.494	
4	9351103	3, 映像の2次情報の論理構造	1.488	○
5	9351103	1, 緒言	1.480	
6	9351103	6, 結言	1.471	
7	9451024	2, 関連研究	1.432	○
8	9351103	5, 実験システム	1.365	○
9	9451024	1, 緒言	1.367	
10	9451024	7, 結言	1.351	

9351103:映像データベースのための論理データモデルとその実装
9451024:映像データベースのための演算体系と圧縮情報構造に関する研究

表1 上位10件のリスト ($\alpha = 0.5, \beta = 0.5$)

	適合率 (precision)	再現率 (recall)
上位3件	0.667	0.200
上位10件	0.500	0.500
上位15件	0.467	0.700

表2 適合率と再現率 ($\alpha = 0.5, \beta = 0.5$)

4. まとめと今後の課題

本研究では，文献が一つの話題だけから成っているのではなく複数の部分的な話題も含んでいることを考え，従来の検索システムより検索精度を高めることのできるような文献検索システムを提案した．

ベクトルを構成する際の重み付けの仕方で文献の長さを考慮するなどの工夫をすれば適合率、再現率が向上すると考えられる．また実装した実験システムの適切な評価の方法を考え，評価したい．

謝辞: 本研究を進めるにあたって日頃から有意義な御指導，御討論をいただいた植村研究室の皆様に感謝いたします．

References

1) Marti A.Hearst. Texttiling:a quantitative approach to discourse segmentation. 1993.

2) M.W. Berry, S.T. Dumais, and T.A. Letsche. Computational methods for intelligent information access. In *Proceedings of Supercomputing*, 1995.

3) Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. In *Journal of the American Society for Information Science*, Vol. 41-6, pp. 391–407, 1990.

4) Christos Faloutsos and Douglas W. Oard. A survey of information retrieval and filtering methods. Technical Report CS-TR-3514, University of Maryland, August 1995.

5) William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval – Data Structures & Algorithms* –. Prentice-Hall, 1992.