知的検索システム Fit での類似度判定の改良*

5N－6

大谷紀子　伊藤史朗　柴田昇吾　上田隆也　池田裕治
キヤノン(株) 情報メディア研究所

1　はじめに

インターネットの普及によってネットワークを介した様々な情報の入手が可能となり，情報洪水が発生している．なかでも，新聞記事のように次々と提供されるフロー情報の利用においては，継続的に新しい情報を収集し，必要な情報を保存して後に活用するという形態が多くとられるため，作業負荷が非常に大きい．そこで，フロー情報の収集から活用までを支援するために，我々は知的検索システム Fit を開発している [1].

Fit では，文書を集めたフォルダをユーザの視点を反映する単位と見なし，フォルダと文書の類似度，或いはフォルダ間，文書間の類似度を判定することで，ユーザの視点に合わせた情報の収集・整理・活用機能を実現している．本稿では，主題を含む段落に着目した類似度判定法とその効果について報告する．

2　Fit の類似度判定

2.1　複数の話題を含む文書

Fit では，フォルダの統合や分割，視点の変化に追従するため，フォルダと文書をベクトルで表現する [2, 3]. 既存文書における単語の出現頻度分布から文書を単位とする統計情報を抽出し，単語の重要度を表す評価値と有効語ベクトルを算出して，ユーザの視点に合ったベクトル空間を形成する．

しかし，文書の各部分が同じ事柄に関する文章とは限らない．導入や例示，比較のために，一文書で複数の事柄を取り上げる場合がある．このように，文書中で述べられている各事柄を話題と呼び，文書の中心となる話題を主題と呼ぶ．複数の話題を含む文書では，付与された視点に合致しない，主題以外の話題の単語がベクトル表現に悪影響を及ぼすと考えられる．

新聞記事等の文書では，内容により文書が段落分けされており，一段落は一話題に対応すると仮定できる．以下では，話題を考慮してベクトル空間を形成するため，段落単位の統計情報を用いた手法について述べる．

*Improvement of Similarity Judgment in Fit, an Intelligent Retrieval System
OTANI Noriko, ITOH Fumiaki, SHIBATA Shogo, UEDA Takaya and IKEDA Yuji (Media Technology Laboratory, Canon Inc.)

2.2　出現段落数に基づく評価値

Fit では，類似度判定における重要度の指標として，各単語の評価値を計算する．評価値は有効語選択の基準となるほか，文書ベクトル算出時の重みとして用いられる．特定のフォルダに偏在する語が類似度判定において重要であると見なし，既存文書における単語の局在度により評価値を算出する．

ところが，文書単位で求めた局在度によると，補足的な話題に関する単語と主題を表す単語の違いが表せず，出現頻度分布が異なっても出現文書数が等しければ評価値が同じ値になる．そこで，出現頻度分布を評価値でより適確に表現するために，有効語候補 T_i のフォルダ F_j に対する帰属度 p_j^i を出現段落数により算出し，これをもとに局在度 $J(T_i)$ を求める．

$$p_j^i = \frac{\text{有効語候補 } T_i \text{ を含むフォルダ } F_j \text{ の段落数}}{\text{フォルダ } F_j \text{ の段落数}}$$

$$J(T_i) = 1 + \sum_{j=1}^{M} p_j^i log_M p_j^i$$

2.3　段落内共起頻度に基づく有効語ベクトル

単語の意味を反映したベクトルで有効語を表現するために，単語間共起関係によるベクトル表現法 [4] を採用している．文書内単語共起を用いることで，同一文書に出現する単語は類似したベクトルで表現され，既存文書における単語出現頻度分布が反映される．

しかし，複数の話題を含む文書では，異なる話題に関連する単語同士の共起も考慮され，各話題の違いを表現できない．そこで，有効語 T_i と基底語 B_j の段落内共起係数 $c_{i,j}$ により有効語ベクトル $\boldsymbol{T_i}$ を定義する．

$$\boldsymbol{T_i} = (c_{i,1}, c_{i,2}, \cdots, c_{i,N})$$

$$c_{i,j} = \frac{T_i \text{ と } B_j \text{ の両方を含む段落数}}{T_i \text{ を含む段落数}}$$

これにより，異なる話題間の単語共起は考慮せず，同一話題における共起のみを考慮することができる．

2.4　他話題段落の除去

各段落に出現する有効語の有効語ベクトルの重み付き平均を段落ベクトルと呼ぶ．理想的なベクトル空間

では，異なる話題に関する段落の段落ベクトルは，互いに離れて位置する．よって，主題以外の話題について述べている段落があると，段落ベクトル平均が主題を表す位置からそれてしまう (図 1(a))．そこで，主題の内容を文書ベクトルで表現するため，主題を表す段落ベクトルのみの平均を文書ベクトルとする (図 1(b))．

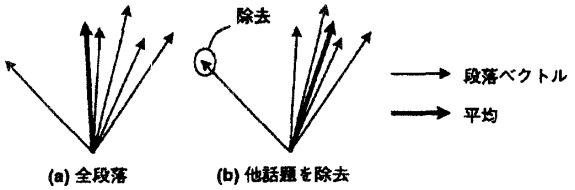

図 1: 段落ベクトル平均

各段落ベクトルが主題を表すか否かは，段落ベクトル間の余弦値の和により判断する．まず，段落ベクトル $\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_N$ について，各段落ベクトルとの余弦値の総和 $COS_1, COS_2, \cdots, COS_N$ を求める．

$$COS_i = \sum_{j=1, j\neq i}^{N} \frac{\boldsymbol{P}_i \cdot \boldsymbol{P}_j}{\|\boldsymbol{P}_i\| \cdot \|\boldsymbol{P}_j\|}$$

$COS_1, COS_2, \cdots, COS_N$ は正規分布をなすと仮定し，平均値 μ，標準偏差 σ により閾値 Th を算出する．

$$Th = \sigma Z + \mu$$

閾値 Th は COS_i の分布の値の小さい方から α% の値であり，Z は α の値に応じて標準正規分布で決定される．COS_i が Th 以上のとき $\boldsymbol{P}_i$ は主題を表し，Th 未満のとき主題以外の話題を表すと判断する．

3 評価

話題の処理の効果を確認するため，新聞記事約 360 記事を用いて実験を行なった．各記事には，10 種類の視点のうち合致する 1〜3 個の視点が人手で付与されている．記事の約半数を学習データとし，残りの半数に対して Fit により視点付与を行なった．閾値設定の基準値を変化させたときの再現率と適合率を図 2に示す．この結果，文書単位よりも段落単位で学習を行なう方が，類似度判定の精度が高いといえる．

また，文書単位の学習と段落単位の学習で形成されるベクトル空間を比較するため，フォルダベクトルの近辺にある有効語ベクトルを調べた．「特許」のフォルダベクトルに近い有効語上位 20 個を表 1に示す．文書単位の学習で挙げられている有効語は，「特許」の特徴を表すとは考えにくいが，段落単位では「著作権」「著作権法」「侵害」等，特許関連の語が多く見られるため，

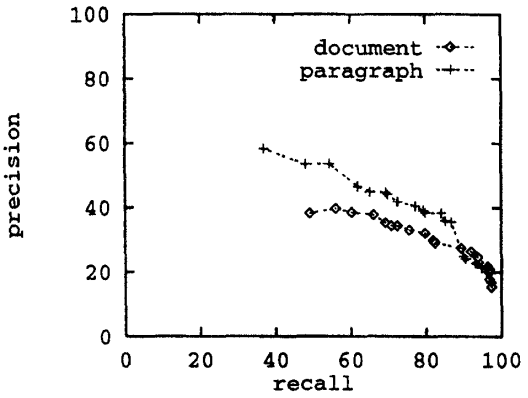

図 2: 再現率・適合率

「特許」という視点の特徴を反映したベクトル空間が形成されたといえる．

表 1: フォルダベクトル近辺の有効語ベクトル

	文書単位		段落単位
1.	マルチメディア	1.	著作権
2.	日本	2.	報告書
3.	現在	3.	著作権法
4.	普及	4.	改正
5.	問題	5.	マルチメディア
6.	利用	6.	方向
7.	技術	7.	デジタル技術
8.	デジタル	8.	デジタル
9.	市場	9.	侵害
10.	時代	10.	範囲
11.	指摘	11.	問題
12.	情報	12.	著作物
13.	パソコン	13.	文化庁
14.	向け	14.	あり方
15.	通信	15.	現在
16.	社長	16.	技術
17.	映像	17.	小委員会
18.	販売	18.	著作権問題
19.	会社	19.	動向
20.	ネットワーク	20.	問題点

4 まとめ

文書で扱われている話題に着目し，段落単位の統計情報が類似度判定に有効であることを確認した．今後は，主題を表す段落のより適切な判別法や，主題以外の段落が有効語ベクトルに及ぼす影響への対処等について検討する．

参考文献

[1] 上田他:フロー情報収集・活用のための知的検索システム Fit(1) コンセプト，第 53 回情処全大, 2T-8, 1996.

[2] 大谷他:フロー情報収集・活用のための知的検索システム Fit(3) 類似度判定，第 53 回情処全大, 2T-10, 1996.

[3] 廣田他:フロー情報を対象にした情報検索システム (4)–文書分類–，第 50 回情処全大, 4F-9, 1995.

[4] 湯浅他:大量文書データ中の単語間共起を利用した文書分類，情処学論, Vol.36, No.8, pp.1819-1827, 1995.