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*

2T－10

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

我々が提案するフロー情報収集・活用のための知的検索システム Fit では，フォルダにより視点を表現し，「視点別の文書提示」，「保存候補のリストアップ」，「フォルダ単位の検索」の各機能を設けている [1]．これらの機能は共通して「類似度判定部」を使用している [2]．本稿では，類似度判定部の処理方式について説明するとともに，類似度判定の性能と，各種文書への適用可能性について評価した結果を報告する．

2　Fit の類似度判定

2.1　処理方式

Fit の類似度判定部では，ベクトル空間モデルに基づいてフォルダや文書を表現し，それらの間の類似度を判定する [3]．各処理について以下で述べる．

有効語辞書の作成

文書のベクトル表現は，文書中の単語の出現パターンに基づく．形態素解析によって既存文書から名詞を抽出し，特定のフォルダに偏在する名詞を有効語とする．この有効語を基底とし，有効語同士の共起確率を成分として，有効語をベクトル表現する．このベクトルを有効語の意味ベクトルと呼ぶ．有効語に関するデータを有効語辞書として保持する．

文書とフォルダのベクトル表現

文書中に含まれている有効語の意味ベクトルの重みつき平均をその文書の文書ベクトルとする．重みは，出現位置と言語的役割をもとに学習により定められる．また，フォルダ中の文書の文書ベクトルの平均をそのフォルダのフォルダベクトルとする．

フォルダベクトルと類似度判定

文書とフォルダ (或いはフォルダ同士) の類似度を判定する時は，文書ベクトルとフォルダベクトルの余弦を計算し，類似度とする．エージェントが視点を付与する際には，視点 V に対応するフォルダ f_V のフォルダベクトル v_{f_V} と到着した文書 d の文書ベクトル v_d の類似度を計算し，その値が設定された閾値以上であるときに，視点 V に合致していると判定する．類似検索時には，類似度をスコアとして，検索されたフォルダをスコアの高い順に並べて提示する．

2.2　フォルダベクトルと有効語辞書の更新

Fit では，時間の経過と共に変化する視点をフォルダの変化によって表現している [1]．フォルダへの文書の追加や削除は，短期的な視点の変化を表すと考えられる．それを類似度判定に反映させるため，フォルダへの文書の追加・削除が行なわれると，その都度フォルダベクトルを更新する．

また，長期的な視点の変化を類似度判定に反映させるために，有効語辞書の更新機能がある．Fit では，有効語辞書のデータに基づいて各種のベクトルを算出するため，このデータにより類似度判定の精度が大きく左右される．初めに作成された有効語辞書は，その時点におけるユーザの視点が反映されているので，視点に沿った類似度判定が行なえる．

しかし，視点は時間の変化と共に変化するので，それに伴うフォルダの分割・統合や新規フォルダの生成に対応するためには，変化した視点に基づいて有効語辞書を作成し直さなければならない．そこで有効語辞書の更新機能により，視点の変化への対応を可能としている．

2.3　類似度の閾値

ある文書が視点に合っているかを判定するときには，あらかじめ視点ごとに設定された類似度の閾値を用いる．ユーザの収集意図を反映した閾値を設定するために，ユーザが収集意図を容易に指示でき，それを適切に閾値で表現する枠組みが必要となる．

類似度が閾値以上のときに視点に合っていると判断するため，閾値を高く設定すると視点に合っている文書を見落としてしまう可能性が高くなり，低い閾値を設定すると視点に合っていない文書までも取り込んでしまう．このような閾値の特性を考慮した上で，「漏れを

*Fit, an Intelligent Retrieval System to Collect and Reuse Flowing Information (3) – Similarity Judgment –
OTANI Noriko, ITOH Fumiaki, SHIBATA Shogo, UEDA Takaya and IKEDA Yuji (Media Technology Laboratory, Canon Inc.)

少なくしたい」「不要文書を減らしたい」などの収集意図を満たすように，閾値を視点ごとに設定しなければならない．

ところが，類似度は文書ベクトルとフォルダベクトルの余弦により表現されており，文書ベクトルの分布範囲の広さは視点ごとに異なるので，ユーザが視点ごとに収集意図を閾値で表現するのは非常に困難である．

そこで，ユーザは期待する再現率を指示することにする．この値を閾値設定の基準値と呼ぶ．閾値設定の基準値を X，フォルダ f_V 中の文書数を N_V，各文書の類似度を $S[1], S[2], \cdots, S[N_V]$ $(\forall i \ \ S[i] > S[i+1])$ とすると，閾値 Th は，

$M = \lfloor N_V * X \rfloor$ （$\lfloor x \rfloor$ は x 以上の最小の整数）

$Th = S[M]$

により求められる．こうすることで，文書ベクトル分布の疎密にかかわらず，視点ごとの収集意図をユーザが自由に表現することができる．

2.4 言語情報を利用した有効語

Fit では，全文書を言語解析して有効語を抽出し，文書やフォルダをベクトル表現している．Fit の処理の中心ともいえる類似度判定の精度を向上させるには，言語解析情報を利用した処理が有効であると考えられる．

そのひとつに複合語の処理がある．抽出した名詞が複合語だった場合，複合語を構成する語基のうち名詞であるものと，複合語自体を有効語の候補にする．これは，語基のみを有効語の候補にするよりも，複合語として意味をなす語の存在により特徴付けられる視点をうまく表現するためである．

例えば，「マルチメディアパソコン」という複合語では，「マルチメディア」と「パソコン」という２つの語基に加えて，「マルチメディアパソコン」という複合語も有効語の候補になる．これにより，マルチメディアパソコンに注目した視点と，マルチメディアに関する視点があった場合，両者を混同することなく，正確に視点付与が行なえるようになる．

複合語の処理について，複合語を構成する語基だけを有効語候補した場合と比べて本方式がどの程度有効かを調べる実験を行なったところ，適合率はほとんど変化しなかったが，再現率が最大5%向上した．

3 評価

類似度判定の精度を評価するため，以下の４種の文書を用いて実験を行なった．

- 新聞記事
- 特許公報
- WWW文書[1]
- 情報検索評価用データベース (BMIR-J1) [2]

上記実験データのうち，新聞記事と特許公報にはそれぞれ10種類，8種類の視点を人手で付与し，WWW文書とBMIR-J1に関しては各文書に付与されている分類名を視点とみなした．実験データの約半数の文書を用いて有効語辞書を作成し，有効語辞書作成に用いなかった文書に対してFitにより視点付与を行なった．閾値設定の基準値が0.6と0.9のときの各視点の平均再現率，平均適合率を表1に示す．

その結果，WWW文書では全体的に再現率が低く，適合率が高いなど，各文書によって再現率，適合率の多少の変動はあるが，どの文書でも十分適用できることが確認された．

表 1: 実験結果

基準値		新聞	特許	WWW	BMIR
0.6	再現率	95.8	80.9	48.7	78.5
	適合率	41.0	50.1	66.2	22.0
0.9	再現率	100.0	92.0	71.0	94.0
	適合率	14.1	27.2	43.5	14.9

4 まとめ

Fit の類似度判定部の処理方式について述べ，類似度判定部の様々な文書への適用可能性を確認した．今後は，言語解析に関するヒューリスティックスの導入などにより，類似度判定の一層の精度向上を目指す．

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，本大会予稿 2T-8, 1996.

[2] 伊藤他:フロー情報収集・活用のための知的検索システム Fit(2) 処理方式，本大会予稿 2T-9, 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–，情報処理学会第 50 回大会 4F-9, 1995.

[1] 日本の様々なWWWページを集めたディレクトリサービス Yahho(http://yahho.ita.tutkie.tut.ac.jp/yahho/) のエンターテイメントに掲載されているページの情報をダウンロードしたもの．

[2] 株式会社 日本経済新聞の協力によって，社団法人 情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが，1993 年 9 月 1 日から 12 月 31 日の日本経済新聞記事を基に構築した情報検索評価用データベース (テスト版)．