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*

2T－9

伊藤史朗　大谷紀子　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]．本稿では，これらを実現するためのシステム構成及び処理方式について説明する．また，フォルダ単位の検索におけるスコア計算の新しい方式を提案する．

2　Fit のシステム構成

Fit では，Boolean モデルとベクトル空間モデルの二つの検索手法を用いている．そのため，文書 d とフォルダ f を，以下のデータ組として保持している．

$$d = (t_d, F_d, v_d) \qquad f = (l_f, c_f, D_f, v_f)$$

ここで，t_dは文書 d のテキスト，F_dは d が属するフォルダのリスト，v_dはテキスト t_dの特徴を表す文書ベクトルである．また，ラベル l_fは利用者がフォルダ fに付ける名称であり，c_fは必要に応じて利用者が設定する検索条件である．D_fは fに保存される文書のリストで，v_fは，fの特徴を表すフォルダベクトルである．

次に，図1に Fit の構成を示し，各部の処理を説明する．

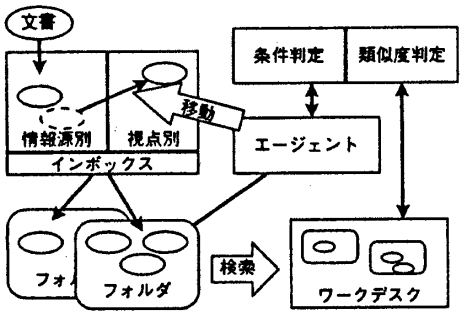

図 1: Fit のシステム構成

条件判定部

Fit で用いられる検索条件 c は，検索語とその論理関係 (AND, OR, NOT) である．条件判定部では，文書のテキスト t_d全文に対して，c が満足されるかを判定する．フォルダに保存されている文書に対しては，保存時に作成されるインデックスを用いる．その他の文書に対しては，パターンマッチングを用いる．

類似度判定部

類似度判定部は，以下の処理を行なう．

- テキストからの文書ベクトルの作成
- 文書ベクトルからのフォルダベクトルの作成
- ベクトルを用いた類似度判定

各処理の詳細は，[2] に述べられている．

エージェント

収集時に他とは分けて見たい視点 Vがあれば，Vに対応するフォルダ f_Vに対して，利用者がエージェントを設定する．エージェントは，Fit に到着する文書 d について，以下の場合に視点 Vに合っていると判定する．

- t_dが c_{f_V}を満足する．
- v_dと v_{f_V}の類似度が設定された閾値を越す．

インボックス

インボックスでは，Fit に到着した文書を利用者に提示する．視点別インボックスと情報源別インボックスとに分かれている．文書はまず情報源別インボックスに入るが，エージェントにより視点 Vに合うと判定された文書は，Vに対応する視点別インボックス b_Vに移る．

b_V中の文書 d をフォルダ f_Vに保存する場合は，保存の指示だけで操作が済む．他の視点のフォルダに保存する場合は，全てのフォルダ fに対し v_fと v_dとの類似度が計算され，類似度の高い順に保存先候補として提示される．利用者は，この中から保存先を選択することができる．

ワークデスク

保存した文書を利用する場合は，ワークデスクでフォルダを検索する．検索されたフォルダに対して，文書の閲覧や移動を行なうことができる．検索手段には，

- 検索条件に合うフォルダを探す条件検索
- 文書 (フォルダ) に似たフォルダを探す類似検索

がある．条件検索においては，次章で説明するスコア付けが行なわれるが，各文書 d がフォルダリスト F_dを保持しているので計算量が軽減される．

3　フォルダの条件検索

3.1　2項母集団の区間推定を用いたスコア付け

文書単位の検索におけるスコア付け方式はいくつか提案されている [3]．しかし，文書ではなくフォルダとして検索条件に合う度合いを示すスコア付けが Fit では必要である．

*Fit, an Intelligent Retrieval System to Collect and Reuse Flowing Information (2) – Retrieval Method –
ITOH Fumiaki, OTANI Noriko, SHIBATA Shogo, UEDA Takaya and IKEDA Yuji (Media Technology Laboratory, Canon Inc.)

そこで，フォルダ f が表現する視点 V_f に合う文書集合 D_V での，検索条件 c を満足する文書の割合 p が，c に対する f のスコア S_fc として適当であると考えた．文書ごとに見ると，視点に合う文書と検索条件 c を満足する文書とは必ずしも一致しない．しかし，文書の集合で見ると，c に合う視点 V ほど p の値は高くなる．

ところで，値 p を直接得ることはできない．そこでまず，観測可能なフォルダ中の文書集合 D_f を用いて，

$$S_fc = \hat{p} = \frac{x}{n}$$

とした．ここで，n は D_f の文書数，x は D_f において c を満足する文書数である．

この $\hat{p}$ は p の最尤推定値になるが，n の値が小さい場合には信頼度が低い．例えば，検索の視点とは無関係でも検索条件 c を満足する文書がありえる．この文書だけから成るフォルダのスコアは 1.0 になり，最も高いスコアとなってしまう．実際の利用では，日々の収集を通してフォルダを形成していくので，文書の少ないフォルダが多くなりやすく，この点は問題である．

そこで，$\hat{p}$ から p の区間推定を行ない，信頼下限 $\hat{p}_L$ をスコアとした．D_V は，割合 p で c を満足する事象を生起する 2 項母集団である．従って，自由度 (ϕ_1, ϕ_2) の F 分布を用いた 2 項母集団の母百分率の区間推定を適用すれば，新たなスコア S'_fc は，

$$S'_fc = \hat{p}_L = \frac{\phi_2}{\phi_1 F^{\phi_2}_{\phi_1}(\alpha) + \phi_2}$$

となる．ここで，自由度は，$\phi_1 = 2(n-x+1), \phi_2 = 2x$ である．また，危険率 2α はパラメータになる．

3.2 フォルダの条件検索の精度評価

提案するスコア付け方式を用いたフォルダの条件検索の精度評価を，情報検索評価用データベース (BMIR-J1) [1] を利用して行なった．

BMIR-J1 は，600 件の新聞記事と 60 件の検索要求文とそれに対する正解集合から成る．まず，各正解集合をそれぞれフォルダとした．次に，どの正解集合にも属さない記事を 1 記事から 5 記事組み合わせてフォルダとした．また，各検索要求を詳細に条件化した条件集合 C_1 と，特徴的と思われる検索語だけで条件化した条件集合 C_2 とを人手で作成した．例えば，「アジア諸国による物資または製品の日本への輸出」に対しては，C_1 の条件は「アジア AND (物資 OR 製品) AND 日本 AND 輸出」であり，C_2 の条件は「アジア AND 輸出」である．なお，C_1 は各々1 条件，C_2 は 3 条件から成る．

これらの検索条件でフォルダを検索し，スコアの高い順にフォルダを並べた．正解のフォルダが検索されればそのフォルダまで，検索されない場合は最後のフォルダまでを採り，それらのフォルダに属する文書を検索結果とみなして再現率と適合率を求めた．なお，比較のために文書単位の検索も行なっている．表 1 に再現率と適合率の平均値を示す．なお，パラメータ α には，実験の結果最も良かった 0.1 を用いている．

表 1: フォルダ検索の精度 [%]

	C_1 による検索		C_2 による検索	
	再現率	適合率	再現率	適合率
文書単位	41.1	57.4	49.7	50.7
S_fc 使用	89.8	72.3	92.8	73.7
S'_fc 使用	89.8	74.6	92.8	76.5

いずれの場合も，区間推定によるスコア付け方式 S'_fc の方が良い結果が得られている．ここで興味深いのは，条件集合 C_1 に比べて C_2 の方が，フォルダ検索の効果が高いことである．厳しい検索条件では，正解フォルダに属する文書でも条件を満足しづらくなるためと考えられる．実際に使う場合，緩い検索条件の方が思い付きやすいので，これは望ましい特徴である．

上記の実験は，検索の視点と各フォルダの視点が一致する場合に相当する．さらに，検索の視点に近い視点しか存在しない場合を評価するため，上記のフォルダ群からランダムに記事を入れ替えたフォルダ群を作成した．このフォルダ群を用いて，条件集合 C_2 について前記と同様に適合率を求めた．ただし，平均再現率が 70% になるまでのフォルダを検索結果として採っている．表 2 に，記事の入れ替え率ごとの平均適合率を示す．

表 2: 文書を入れ替えた場合の平均適合率 [%]

入れ替え率	10%	20%	30%	40%	50%
S_fc 使用	68.7	65.7	60.5	42.5	34.3
S'_fc 使用	72.8	67.6	62.3	43.6	35.5

入れ替え率が 30% 程度までは，フォルダ検索並びに区間推定によるスコア付けの効果は高いが，それ以上では文書単位で検索した方が良い．しかし，実際のフォルダは入れ替え率 30% 以内に相当すると考えており，利用実験でもフォルダ検索の有効性が確認されている [4].

4 まとめ

以上，Fit の処理方式について述べた．また，Fit で使用しているフォルダ検索の新しいスコア付け方式の効果を確認した．

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] 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，本大会予稿 2T-10, 1996.
[3] 高木他:単語出現共起関係を用いた文書重要度付与の検討，情処研報，96-FI-41, 61-68, 1996.
[4] 伊藤他:フロー情報の収集・活用を支援する検索システム，情処研報，96-DD-1, 15-22, 1996.

[1] 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).