ユーザへの適応性を考慮した適合フィードバックによる WWW 情報検索

非会員 江 口 浩 二 (関西大)

正 員 伊 藤 秀 隆 (関西大)

正 員 隈 元 昭 (関西大)

Relevance-Feedback-Based WWW Information Retrieval Considering Adaptation to User's Behavior
Koji EGUCHI, non member, Hidetaka ITO, member, Akira KUMAMOTO, member (Kansai University)

The recent growth of information resources on the WWW has led to the increasing need for developing flexible methods of information retrieval. Since the behaviors of users of WWW information retrieval systems naturally reflect the complex hypertext structure of the WWW pages written from various viewpoints, it is desirable that the systems should be able to track various dynamic changes in the intentions of the users. Most of conventional techniques of information retrieval, which typically assume that the users have precise and fixed goals, are not suitable for realizing the desired features for WWW information retrieval systems. In this paper, we propose a new extension to the relevance feedback and demonstrate its adaptability to a wide variety of the behavior of users.

キーワード: WWW, 情報検索, 適合フィードバック, ユーザモデル

1. まえがき

近年, 情報スーパーハイウェイ⁽¹⁾の構築, 電子図書館⁽²⁾の研究開発や World Wide Web(WWW)の発展にともない, より柔軟な情報検索手法への要求が高まっている。

WWW は, 膨大な量の情報資源から成るハイパーテキスト構造のネットワークであり, それぞれの情報資源は, HTML(Hypertext Markup Language)⁽³⁾により半構造化されている。また, WWW ページ(HTML 文書)は多様な観点により記述されている。このような WWW の特徴を反映して, WWW 情報検索においてはユーザの検索目標が多様かつ動的に変化するものと思われ, システムがこれに追従することが望ましい。

さて, 従来の情報検索手法のほとんどは, ユーザが明確な検索目標を有する場合に焦点を当てているものと思われる。西村ら⁽⁴⁾は, キーワードが検索質問として同時に使用される傾向から, キーワード間の相関性を導出して, システムがそれに基づいて関連キーワードを提示することにより, ユーザを支援する手法を提案している。このような手法は, ユーザが検索開始時には漠然とした検索目標しか有さないが検索過程においてユーザの興味が絞り込まれ, 次第に具体的な検索目標を設定していくような検索行為を支援する目的にも有効であろう。

本論文では, これとは異なるアプローチにより, ユーザの多様な検索行為を支援するため, 伝統的な適合フィードバック¹(Relevance Feedback)⁽⁵⁾を拡張し, WWW 情報検索に適用する^(6,7)。提案する手法は, ユーザとのインタラクションを通じてユーザの興味の変化を検出し, それに応じて, 動的にシステムの状態を変化させ, 時間的に変化するユーザの検索目標に適応するものである。

2. 提案手法の枠組み

提案手法の枠組みを図 1 に示す。提案手法の構成要素それぞれについて, 以下に説明する。

<2.1> ロボット WWW は, 膨大な量の情報資源から成るハイパーテキスト構造のネットワークであるが, このような情報空間を対象とする場合, 何らかの形で事前に情報資源の各々のインデックスを作成しておくことが必要となる。本論文では, これを実現するために, ロボット⁽⁸⁾と呼ばれるソフトウェアモジュールを用いる。

本論文では, WWW の情報資源の中で特に重要な HTML 文書ファイルに着目する。ロボットは, 上記のようなハイパーテキスト構造を有する WWW 情報空間の特性を利用し, HTML 文書を取得するとともに HTML

¹ 関連フィードバック, 関連性フィードバックとも呼ばれる。

図 1 提案手法の枠組み

Fig. 1. Framework of the proposed method.

文書内のリンクを抽出し、更に、HTML 文書に対してキーワード抽出などのインデクシングを行う。次に、ロボットは抽出されたリンクを辿って、新たな HTML 文書を取得する。これを再帰的に繰り返すことにより、情報資源を収集する。最終的に、情報空間の特徴を写像した、インデックス・データベースを生成する。インデックス・データベースは各情報資源のキーワード群の他に、URL(Uniform Resource Locators)⁽³⁾、タイトル、ファイルサイズ、最終更新日等の情報を格納する。

本論文では、HTML 文書のそれぞれについて、インデックスとしてのキーワードの表現にベクトル空間モデル⁽⁹⁾を用いる。ここで、文書ベクトルの基底は HTML 文書から抽出されたキーワードであり、文書ベクトルの成分は、その重要度を表す重みである。

<2.2> インデクシング 妥当なキーワードを切り出すために、HTML 文書の特徴づけていると思われる、以下のような HTML タグに着目する。

<title>, <h1>, <h2>, <h3>, <h4>, <a>,
, , ,

これらのタグで特徴づけられた文字列に対して、形態素解析²を行い、名詞と未定義語³をキーワード候補とする。これらに対して、重みづけを行うが、本論文では以下の TF-IDF 法を用いる^(11,9)。TF-IDF 法は通常の情報検索において用いられてきたものであるが、WWW 情報検索においてもその有効性が確認されている⁽¹²⁾。

² 日本語形態素解析システムとして、『茶筌』⁽¹⁰⁾を用いた。

³ 検索効率のため、平仮名のみからなる語、数字、特殊文字を除き、片仮名は 2 文字以上、アルファベットは 3 文字以上とする。また、HTML 文書特有の HTML タグ、E メールアドレス、URL 等を除く。更に、キーワードとして不適切な、高頻度語等のストップワードを除く。

- (1) 文書 d において語 t の出現頻度を tf_t^d とする。
- (2) 文書の総数を N 、語 t を含む文書の数を df_t として、 $idf_t = \log(N/df_t)$ により、 idf_t を求める。
- (3) f を (1) で求めた tf_t^d を引数とする単調増加関数であるとして、文書 d における語 t の重み w_t^d を次式により求める。

$$w_t^d = f(tf_t^d) \cdot idf_t \quad \dots\dots\dots (1)$$

<2.3> 類似度計算 ユーザの入力した 1 語以上のキーワードを、<2.1>および<2.2>節で述べた、文書を表記するのに用いたベクトル空間と同一の空間上に表現する。ただし、正規化を行なう。そして、検索質問ベクトル q と文書ベクトル d_r の内積を類似度とする。即ち、 $\text{sim}(q, d_r) = (q, d_r)$ である。類似度の大きいものから順位づけしたものを検索結果とする。

<2.4> 適合フィードバック 情報検索における効率的な学習手法として適合フィードバックが知られている。これは、検索結果に対してユーザが行った適合、不適合の評価を、検索質問を成す個々のキーワードの重みに寄与させ、検索精度を高めようとするものである。ここでは、Rocchio の式⁽⁵⁾を用いて、これを実現する。

$$q_{k+1} = \hat{q}_k + \frac{\alpha}{|R|} \sum_{\hat{d}_i \in R} \hat{d}_i - \frac{\beta}{|N|} \sum_{\hat{d}_j \in N} \hat{d}_j \quad \dots\dots\dots (2)$$

ただし、 α, β はフィードバックの度合いを示すパラメータ、 R, N はそれぞれ検索結果に対する適合文書の集合、不適合であるとわかっている文書の集合を表す⁽¹³⁾。このように検索結果に対して適合とも不適合とも判断できない場合を許し、この場合の文書に関しては適合フィードバックの対象としない。また、 q_{k+1} において、重みが負の値をとる語は除去する。なお、ベクトル $\hat{q}_k, \hat{d}_i, \hat{d}_j$ はそれぞれ正規化されているものとする。

ところで、式 (2) において、パラメータ α, β の値については、それぞれ 2, 0.5 が多く用いられている^(14,13)。パラメータ α, β を動的に変化させる手法として、Buckley ら⁽¹⁵⁾は TREC-3⁽¹⁶⁾において、 α, β を変化させて、再検索を数回繰り返し、予め与えられた学習文書集合 (learning document set) の検索精度が良好となる結果を用いるという方法をとっているが、WWW 情報検索のように実時間性が要求される目的には、相応しくないとと思われる。

<2.5> 適合フィードバックの動的パラメータ調整 本論文では、新たなアプローチとして、ユーザが評価した適合/不適合文書と検索質問との距離に基づいて α, β の値を動的に変化させ、ユーザの多様な検索行為に対処することを試みる。

さて、ユーザの多様な検索行為は、次のように大別できると考える。

- (i) ユーザが明確な検索目標を有し、ある程度それに即した検索条件を記述して、検索を行う場合

(ii) ユーザが検索開始時には漠然とした検索目標しか有さないが、検索過程においてユーザの興味が絞り込まれ、漸次的に具体的な検索目標を設定していく場合

(iii) ユーザが検索過程において新たな知識を発見し、新たな検索目標を設定する場合

以上のような、ユーザの多様な検索行為に対応するため、次の仮定を設ける。まず、適合評価について、

(1) 検索質問と適合文書の距離が近接していない場合として、例えば、(iii) のようにユーザの翻意が発生した場合が考えられる。このとき、式(2)の α は大きくとり、適合評価した文書から得られる情報を特に強調する。

(2) 検索質問と適合文書の距離が近接している場合として、例えば、(i) のようにユーザが明確な検索目標を有し、検索精度の向上を期待している場合が考えられる。このとき、式(2)の α の値は従来用いられてきた 2 に近い値をとる。

不適合評価については、これとは逆に、

(1) 検索質問と不適合文書が近接していない場合、式(2)の β の値は従来用いられてきた 0.5 に近い値をとる。

(2) 検索質問と不適合文書が近接している場合、式(2)の β は大きくとり、不適合評価した文書から得られる情報に関して検索質問から重みを特に削減する。

以上のような考えのもと、パラメータ α, β を検索質問ベクトルと文書ベクトルの内積の最大値により求める。すなわち、パラメータ α, β を次式のように求める。

$$\alpha = \begin{cases} 1 / (x_\alpha + y_\alpha \cdot \max_{d_i \in R} \langle \hat{q}_k, \hat{d}_i \rangle) & (\max_{d_i \in R} \langle \hat{q}_k, \hat{d}_i \rangle \leq a) \\ 2 & (\max_{d_i \in R} \langle \hat{q}_k, \hat{d}_i \rangle > a) \end{cases}, \dots \dots (3)$$

$$\beta = \begin{cases} 0.5 & (\max_{d_j \in N} \langle \hat{q}_k, \hat{d}_j \rangle \leq b) \\ x_\beta + y_\beta \cdot \max_{d_j \in N} \langle \hat{q}_k, \hat{d}_j \rangle & (\max_{d_j \in N} \langle \hat{q}_k, \hat{d}_j \rangle > b) \end{cases} \dots \dots (4)$$

ただし、 $\hat{d}_i, \hat{d}_j$ および $\hat{q}_k$ は正規化されている。また、 $\max_{d_i \in R} \langle \hat{q}_k, \hat{d}_i \rangle = a$, $\max_{d_j \in N} \langle \hat{q}_k, \hat{d}_j \rangle = b$ において、関数が連続となるように、定数 $x_\alpha, y_\alpha, x_\beta, y_\beta$ を決定する。

3. システムの実装および実験

<3.1> 対象領域 関東、関西圏の大学情報工学系53学科の、日本語で記述された HTML 文書を約1万件収集し、WWW上に提案手法のプロトタイプを実現した。大学情報工学系学科の HTML 文書に限定することで、十分に構造化された HTML 文書を収集することができ、なおかつ、大量で多様な WWW 情報空間の特性を十分に反映できると考える。

<3.2> インタフェース プロトタイプシステムのインタフェースを図2に示す。ユーザは、まず検索質問をシステムに入力する。これに対して、システムは検索質問とインデックス・データベースに格納された文

図2 プロトタイプシステムのインタフェース

Fig. 2. Interface of the prototype system.

書とを照合して、タイトル、キーワード、URL、ファイルサイズ、最終更新日、類似度、ユーザ評価用のボタンを1セットとした検索結果を表示する。ここで、タイトル、URLは実際のWWWページへとリンクされている。ユーザは、検索結果に付与されるそれぞれのWWWページのキーワードからページの内容を推測するか、あるいは、実際にページをブラウジングすることによって、適合、不適合、わからないの評価を行う。このとき、それぞれの評価に応じて、「Good」、「NG」、「?」のいずれかのボタンを選択する。「Feedback」のボタンをクリックすることでユーザの評価を考慮した再検索が行われる。これを数回繰り返すことによって、ユーザの検索目標に合致した検索結果が得られていく。なお、「Contents of Query」のボタンをクリックすると、キーワードとその重要度からなる検索質問の内容が、別枠のウィンドウに表示される。これにより、ユーザは検索時の詳細な検索条件を知ることができる。

<3.3> インデクシング 式(1)として、次の3種の式を考える。

$$w_t^d = tf_t^d \cdot idf_t, \dots \dots \dots (5)$$

$$w_t^d = \left[0.5 + 0.5 \frac{tf_t^d}{\max_{t \in T^d} tf_t^d} \right] \cdot idf_t, \dots \dots \dots (6)$$

$$w_t^d = [\log(tf_t^d) + 1.0] \cdot idf_t, \dots \dots \dots (7)$$

ただし、 T^d は文書 d におけるキーワード候補集合である。

ところで、式 (5),(6),(7) はいずれも通常の情報検索において用いられてきた^(11,14)ものであるが、通常の情報検索では、 $\hat{w}_i^d = w_i^d / \sum_{i \in T^d} w_i^d$ のように正規化を行うことが多い。しかしながら、WWW のようなハイパーテキストでは、個々のページは短い文章で表現される傾向があり、このような場合、正規化を行わない方が精度が良いことが経験的にわかっている⁽¹²⁾。本論文においても文書ベクトルの正規化を行わない。

計算コストを考慮して、HTML 文書の 200 行目までを対象にキーワード抽出を行い、得られたキーワード候補に対して、式 (5),(6),(7) で表される関数を用いて異なる重みづけを行い、それぞれの精度を比較する。

本論文では、ベクトル空間の次元数は約 25000 であり、インデクシングにおいて、文書ベクトルは重みの順に 10 のキーワードのみについてその重みをベクトルの成分とし、その他のベクトルの成分を 0 とした。

検索精度の評価には、情報検索で一般的に用いられる適合率、再現率を用いた⁽¹⁷⁾。すなわち、

適合率 = $\frac{\text{検索された適合文書数}}{\text{検索された文書の総数}}$, (8)

再現率 = $\frac{\text{検索された適合文書数}}{\text{適合文書の総数}}$ (9)

適合または不適合の評価は本来主観的であり、ユーザによって異なるものである。ここでは、第一著者が一定の判断基準⁴をもって実際の WWW ページの内容に基づいて評価した。また、式 (9) の適合文書の総数は、3 種の重みづけ関数で検索された適合文書の和集合で近似した。

実験のため、表 1 のような情報工学系分野の 10 のトピックスを初期検索質問として、これらの検索精度の平均をとり、このときの再現率・適合率グラフを図 3 に示す。また、再現率が 0.0, 0.1, ..., 1.0 の 11 点における適合率の平均値を平均適合率とし、表 2 に示す。平均適合率は、検索精度の評価のため多く用いられている。

表 1 初期検索質問
Table 1. Initial queries.

	検索質問		検索質問
1	Data_Mining	6	遺伝アルゴリズム
2	マルチエージェント	7	プログラミング
3	Neural_Network	8	ヒューマンインタフェース
4	推論	9	視覚化_Visualization
5	ロボット	10	マルチメディア

図 3 および表 2 より、3 種の重みづけ関数の精度はそれほど差異が現れなかった。ところで、式 (6) の重みづけ関数は tf_i^d の最大値の計算にコストを要求するため、

⁴大学の講義のシラバス、研究室名のリスト等是不適合とした。

図 3 重みづけ関数の比較 (再現率・適合率グラフ)
Fig. 3. Comparison of three weighting methods (recall-precision graph).

表 2 重みづけ関数の比較 (平均適合率)
Table 2. Comparison of three weighting methods (average precision).

式 (5)	式 (6)	式 (7)
0.680	0.659	0.715

WWW のように大量のデータを扱う上で好ましくない。式 (5),(7) のうち、本論文ではこれ以後、式 (7) の重みづけ関数^(18,14)を用いることにする。

<3.4> 適合フィードバック 情報検索の分野において研究されてきた適合フィードバックを、WWW 情報検索に適用した研究⁽¹⁹⁾はあるが、その効果について充分検討されていないようである。

適合フィードバックの効果を確認するため、表 1 の検索において、適合文書数が比較的多かった検索質問 2, 3, 5, 8, 10 の 5 つについて、式 (2) における α, β の値をそれぞれ 2, 0.5 に固定して、適合フィードバックにより再検索を行った。このときの検索精度の平均を求め、再現率・適合率グラフを図 4 に、平均適合率を表 3 (パラメータ固定の場合) に示す。ただし、フィードバック時の適合/不適合評価は検索結果の上位 20 件のみについて行い、また、検索精度は上位 60 位以内について求め、61 位以下は無視した。式 (9) における適合文書の総数は、パラメータ固定の場合に検索された適合文書と<3.5.1>項で調整を行った場合のそれとの和集合で近似した。図 4 と表 3 (パラメータ固定の場合) より、従来、情報検索で用いられてきた適合フィードバックが、WWW 情報検索においても検索精度の向上に有効であるといえる。

<3.5> 適合フィードバックの動的パラメータ調整 <2.5> 節で述べた適合フィードバックの動的パラメータ調整を実現するため、式 (3),(4) において $a = 0.679$, $b = 0.339$ とした。これらは、表 1 の検索質問について、一定の検索目標をもってフィードバックを 3 度繰り返したときの、式 (3),(4) に示した $\max_{d_i \in R} (\hat{q}_k, \hat{d}_i)$,

図4 適合フィードバックの効果 (再現率・適合率グラフ)
Fig. 4. Effect of the relevance feedback (recall-precision graph).

$\max_{d_j \in N} \langle \hat{q}_k, \hat{d}_j \rangle$ のそれぞれの平均値であり、 a については検索質問ベクトルと文書ベクトルの距離が充分近くなっている指標、 b については充分距離がある指標であるとする。

また、 $x_\alpha, y_\alpha, x_\beta, y_\beta$ の値については、 $(x_\alpha, y_\alpha) = (0.010, 0.722)$ 、 $(x_\beta, y_\beta) = (0.244, 0.756)$ とした。このときの α, β の特性を図5に示す。図5のように、ユーザの

図5 適合フィードバックの係数の特性
Fig. 5. Characteristics of coefficients in the relevance feedback.

検索目標の変化に適応させるために、主に式(2)における α の値を変化させ、 β の値は大きくは変化させない。これは、 β の値を大きくすると、かえって検索効率が低下する場合があるためである。

<3.5.1> 明確な検索目標を有するユーザモデル
<2.5>節に述べた、明確な検索目標を有し、ある程度それに即した検索条件を記述して検索を行うユーザモデル(i)について、実験を行う。<3.4>節の実験と同様に、表1の検索質問のうち、適合文書数の比較的多い5つについて、提案手法を用いて、初期検索、1度目、2度目および

3度目の再検索を行い、それぞれに5つの検索質問についての検索精度の平均をとる。更に平均適合率を求め、表3(パラメータ調整の場合)に示す。なお、5つの検索質問に関する α, β の平均値を $\bar{\alpha}, \bar{\beta}$ として、表3に付記する。また、2度目の再検索を行ったときの再現率・適合率グラフを図6に示す。

表3 明確な検索目標を有するユーザモデルについての実験結果 (平均適合率の比較)

Table 3. Experimental results on a user-model which has precise goal in retrieval (average precision).

	パラメータ 固定の場合	パラメータ調整の場合		
	平均適合率	平均適合率	$\bar{\alpha}$	$\bar{\beta}$
初期検索	0.518	0.518	(2.00)	(0.500)
再検索 (1)	0.730	0.736	2.51	0.661
再検索 (2)	0.750	0.767	2.15	0.613
再検索 (3)	0.855	0.854	2.12	0.589

図6 明確な検索目標を有するユーザモデルについての実験結果 (再現率・適合率グラフ)

Fig. 6. Experimental results on a user-model which has precise goal in retrieval (recall-precision graph).

表3および図6より、本ユーザモデルでは、適合フィードバックにおいて従来のパラメータを固定する方法とほぼ同様な検索精度が提案手法でも得られることがわかる。これより、ユーザが明確な目標を持つ場合については、提案手法を用いることで、検索精度の向上は認められないものの、従来の固定のパラメータを用いる適合フィードバックの検索精度を損なわないことが確認される。

<3.5.2> 検索目標の翻意を行うユーザモデル
<2.5>節に述べたような、検索過程において新たな知識を発見し新たな検索目標を設定するユーザモデル(iii)について、式(2)のパラメータ α, β を固定した場合の検索精度と、動的に調整を行った場合のそれとを実験により比較する。

表1の検索質問のうち、適合文書数の比較的多かった検索質問2, 3, 5, 8, 10について、表4に示すような検索目標の翻意を想定して以下に述べるような手順で実験する。

表 4 検索目標の変化

Table 4. Changing retrieval goals of users.

	前回の検索目標	新たな検索目標
2	マルチエージェント	オブジェクト指向
3	ニューラル ネットワーク	ロボティクス
5	ロボティクス	画像認識・理解
8	ヒューマン インタフェース	対話処理
10	マルチメディア	画像認識・理解

検索質問 5 を例にあげる。まず、「ロボティクス」について検索し、適合フィードバックによる再検索を 1 度行う⁵。次に、検索結果の中で「画像認識・理解」に関するものを適合と評価し、式 (2) におけるパラメータ α が固定の場合の検索精度と、動的に調整を行った場合とを同一の条件で比較する。調整を行った場合の α の値は 6.10 であった。ただし、簡単のため不適合評価は行わず、検索精度の計算において検索結果の上位 61 件以降は無視し、式 (9) の再現率の計算に必要な適合文書の総数は本項の実験において検索された適合文書の和集合で近似する。

上記 5 つの検索目標の翻意例について検索精度の平均をとり、再現率・適合率グラフを図 7 に示す。平均適合率

図 7 翻意を行うユーザモデルについての実験結果 (再現率・適合率グラフ)

Fig. 7. Experimental results on a user-model which revises goal in retrieval (recall-precision graph).

はパラメータ固定の場合 0.465、動的調整を行った場合 0.603 であった。なお、調整を行ったときの α の平均値は 11.5 であった。図 7 および平均適合率より、パラメータを固定にしたときよりも動的調整を行ったときの方が、ユーザの翻意に追従できることが確認される。

<3.5.3> 興味の絞り込みを行うユーザモデル
<2.5>節に述べた、ユーザが検索開始時には漠然とした検索要求しか有さないが検索過程においてユーザの

⁵式 (3),(4) によりパラメータの動的調整を行う。このとき、 $(\alpha, \beta) = (2.95, 0.640)$ であった。

興味が絞り込まれるようなユーザモデル (ii) について考察する。

ユーザが興味を絞り込み、検索結果に対してそれに即した適合／不適合評価を行うとき、検索質問と適合文書の距離が近接しているならば、その際の検索条件は大きく修正されなくとも良いと考えられる。このとき、<3.5.1>項の明確な検索目標を有するユーザモデルと同様に、式 (2) における α の値は式 (3) に基づいて 2 に近い値となる。

逆に、検索質問と適合文書が近接していないならば、検索条件は適合文書のキーワードをもとに大きく修正されるべきで、<3.5.2>項の翻意を行うユーザモデルと同様、式 (2) における α の値は式 (3) に基づいて大きくなる。

以上のように、興味の絞り込みを行うユーザモデルに関しては、<3.5.1>項の明確な検索目標を有するユーザモデルかあるいは<3.5.2>項の翻意を行うユーザモデルと同様に対処されるが、これは妥当な処置であると考えられる。

4. むすび

本論文では、適合フィードバックが WWW 情報検索にも有効であることを確認し、更に、情報検索における、ユーザの多様な検索行為に対応するため、適合フィードバックのパラメータを検索質問と文書の距離に基づいて動的に調整する手法を提案し、WWW 情報検索に適用した。実験の結果、ユーザの検索目標が多様かつ動的に変化する場合およびユーザが漠然とした検索目標を有する場合について提案手法の有効性を確認した。また、ユーザが明確な目標を持つ場合についても、提案手法を用いることで、検索精度の向上は認められないものの、従来の適合フィードバックの検索精度を損なわないことを確認した。しかしながら、以下の問題点があげられる。

- (1) WWW 情報検索に適合フィードバックを用いるにあたって、ユーザに与える負荷を考慮すると、現実的には検索結果の上位を占めるたかだか 20 件程度に対して、ユーザが適合／不適合評価を行うことになる。しかしながら、同様な特徴を持つ文書が上位を占める場合や、適合文書数が極端に少ない場合、それらに対するユーザの評価をフィードバックしても、検索結果は局所的に収束するのみであり、発見的な情報を得ることが容易でない。
- (2) ユーザに対して、検索結果から適合／不適合の評価を求める際に、ユーザへ与える負荷が大きい。

以上の問題に対処するため、検索結果をクラスタリング⁽²⁰⁾、その文書空間を計算機の画面上に表現して、インタフェースとすることが考えられる。これにより、多くの WWW ページに対して適合／不適合評価を行うことができ、検索結果の早熟な収束を避けられると考える。また、検索結果のクラスタリングを視覚化することにより、ユーザが多くの検索結果の中から、文書クラス間の相対的な距離、あるいは、質問・文書クラス間の距離を

手掛かりに、適合／不適合文書を評価することを支援することが考えられる。また、最近活発に研究されつつある情報抽出技術⁽²¹⁾を用いて、WWW ページの適合／不適合評価を容易にすることも検討に値する。

以上のように、本論文で提案した動的パラメータ調整機能を有する適合フィードバックを、検索結果のクラスタリング、情報視覚化、情報抽出技術などと組み合わせることが今後の課題である。

謝辞 本研究において、ご協力下さった関西大学工学部 藤本剛司氏、同博士前期課程山口洋介氏に感謝します。

(平成9年3月4日受付, 平成9年8月25日再受付)

文 献

- (1)“特集:情報ハイウェイの展開と利用”, 信学誌, **78**, 4, pp. 333-426 (1995).
- (2)“特集:デジタル図書館”, 情報処理, **37**, 9, pp. 813-864 (1996).
- (3)The World Wide Web Consortium, : “W3C — The World Wide Web Consortium”, <http://www.w3.org/> (1996).
- (4)西村英樹, 伊藤耕一郎, 河野浩之, 長谷川利治: “重み付き相関ルール導出アルゴリズムによる WWW データ資源の発見”, 電子情報通信学会 第7回データ工学ワークショップ (DEWS'96), pp. 79-84 (1996).
- (5)Rocchio, J. J.: “Relevance feedback in information retrieval”, The SMART Retrieval System: Experiments in Automatic Document Processing (Ed. by Salton, G.), Prentice Hall, pp. 313-323 (1971).
- (6)江口浩二, 山口洋介, 伊藤秀隆, 隈元昭: “WWW における知的情報検索の一提案”, 平成8年電気関係学会関西支部連合大会, G14-7, p. G343 (1996).
- (7)江口浩二, 藤本剛司, 山口洋介, 伊藤秀隆, 隈元昭: “ユーザの翻意を考慮した適合フィードバックによる WWW 情報検索”, 1997年信学総大, D-4-1, p. 74 (1997).
- (8)Koster, M.: “World Wide Web robots, wanderers, and spiders”, <http://info.webcrawler.com/mak/projects/robots/robots.html> (1996).
- (9)Salton, G.: “Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer”, Addison-Wesley (1989).
- (10)松本裕治, 北内啓, 山下達雄, 今一修, 今村友明: “日本語形態素解析システム『茶筌』version 1.0 使用説明書” (1997).
- (11)Salton, G. and Buckley, C.: “Term-weighting approaches in automatic text retrieval”, Information Processing & Management, **24**, 5, pp. 513-523 (1988).
- (12)Yuwono, B. and Lee, D. L.: “Search and ranking algorithms for locating resources on the World Wide Web”, Proceedings of the 12th International Conference on Data Engineering, pp. 164-171 (1996).
- (13)Buckley, C., Salton, G. and Allan, J.: “The effect of adding relevance information in a relevance feedback environment”, Proc. ACM SIGIR '94, pp. 292-298 (1994).
- (14)Buckley, C., Allan, J. and Salton, G.: “Automatic routing and ad-hoc retrieval using SMART: TREC 2”, Proceedings of the Second Text REtrieval Conference (TREC-2) (Ed. by Harman, D. K.), NIST Special Publication 500-215, pp. 45-56 (1994).
- (15)Buckley, C. and Salton, G.: “Optimization of relevance feedback weights”, Proc. ACM SIGIR '95, pp. 351-357 (1995).
- (16)Harman, D. K.: “Overview of the third Text REtrieval Conference (TREC-3)”, Proceedings of the Third Text REtrieval Conference (TREC-3) (Ed. by Harman, D. K.), NIST Special Publication 500-225, pp. 1-19 (1995).
- (17)伊藤哲朗: “情報検索”, ソフトウェア講座 19, 昭晃堂 (1986).
- (18)Buckley, C. and Allan, G. S. J.: “Automatic retrieval with locality information using SMART”, Proceedings of the First Text REtrieval Conference (TREC-1) (Ed. by Harman, D. K.), NIST Special Publication 500-207, pp. 59-72 (1993).
- (19)Yuwono, B. and Lee, D. L.: “WISE: A world wide web resource database system”, IEEE Transactions on Knowledge and Data Engineering, Special Section on Digital Library, **8**, 4, pp. 548-554 (1996).
- (20)Hearst, M. A., Karger, D. and Pedersen, J. O.: “Scatter/Gather as a tool for navigation of retrieval results”, 1995 AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval, pp. 65-71 (1995).
- (21)武田英明: “ネットワークを利用した知的情報統合”, 人工知能学会誌, **11**, 5, pp. 680-688 (1996).

江口 浩二

(非会員) 1993年同志社大学工学部電子工学科卒業。1995年関西大学大学院工学研究科博士前期課程修了。現在、同博士後期課程在学中。情報検索, Human-Computer Interaction, 自然言語処理などに関心を持つ。情報処理学会, 電子情報通信学会各会員。

伊藤 秀隆

(正員) 1994年京都大学大学院博士後期課程修了。同年、関西大学工学部助手。京都大学博士(工学)。非線形力学現象, 知識情報処理などに関する研究に従事。電子情報通信学会, 人工知能学会, 計測自動制御学会各会員。

隈元 昭

(正員) 1971年京都大学大学院修士課程修了。同年、関西大学工学部助手。現在、同助教授。京都大学工学博士。知識情報処理, 計算機ソフトウェア, 画像処理, CADなどの研究に従事。電子情報通信学会, システム制御情報学会, 計測自動制御学会各会員。