

ユーザへの適応性を考慮した WWW 情報検索における漸次的なクエリの拡張

江口 浩二[†] 伊藤 秀隆[†] 隈元 昭[†]

[†]関西大学 工学部 電気工学科

〒 564 吹田市山手町 3-3-35

e-mail: eguchi@enzan.ee.kansai-u.ac.jp

あらまし 近年, World Wide Web(WWW) の普及や電子図書館の構築に伴い, ユーザフレンドリな情報検索インターフェースへの要求が高まっている。特に, WWW 上の HTML 文書は多様な観点で記述されているため, それらの検索においてはユーザの検索目標が漠然としていたり, 動的に変化することも多いと考えられ, それに対処するための適合フィードバックとその拡張手法が検討されている。本稿では, これらの手法を更に有効に機能させるため, (1) ユーザの興味を反映した対話的な文書クラスタリング, (2) 動的に変化する検索目標を反映した漸次的なクエリの修正とユーザへの提示を提案する。

Incremental Query Expansion Considering Adaptation to User's Behavior on the WWW

Koji EGUCHI[†] Hidetaka ITO[†] Akira KUMAMOTO[†]

[†]Department of Electrical Engineering, Faculty of Engineering, Kansai University
3-3-35 Yamate-chou, Suita, Osaka, 564 Japan
e-mail: eguchi@enzan.ee.kansai-u.ac.jp

Abstract The recent growth of the World Wide Web (WWW) and digital libraries has led to the increasing need for developing user-friendly human interfaces for information retrieval systems. In particular, the intentions of the users retrieving HTML documents on the WWW, which are written from various viewpoints, are often vague and change as the retrieval proceeds. For such cases, several attempts to apply the relevance feedback and its extensions have been carried out with some effectiveness observed. To enhance the effectiveness of this framework, we propose in this paper (1) interactive document clustering reflecting interests of the users and (2) incremental query expansion reflecting the changing intentions of the users.

1 はじめに

近年, インターネットの普及とともに膨大な情報にアクセスできる環境が提供されつつあり, 特に WWW の普及は目覚ましい。それに伴い, WWW ベースの電子図書館の研究開発が活発になりつつある一方で, CALS(Commerce At Light Speed) の標準文書形式として半構造化文書形式 SGML が採用され, 技術文書の共有化が進められている。

これらの要素技術として, 複数の機関によってインターネット上で公開されている広域に分散した情報資源 (SGML 文書等) から検索に必要なインデックス情報を自動抽出することにより, それらの情報資源を横断的に

検索し, アクセスする技術が必要となってきた。このとき, 多種多様な情報資源から必要な情報を的確に見い出す作業は, ユーザに熟練した経験および知識を要求するため, このような, ユーザに課せられた負荷を軽減することが望まれる。また, 特に WWW の情報検索においては, ユーザの興味が漠然としていたり, ユーザの検索目標が動的に変化すること等が顕著であると考えられ, システムがこれに適応することが望ましい。

以上のような問題意識から, 我々は, 伝統的な適合フィードバック¹(Relevance Feedback)[1] を拡張し, 動的に変化するユーザの検索目標への適応を目指す手法[2, 3, 4](以下, 拡張適合フィードバック)を提案し, WWW

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。

情報検索への適用を検討してきた。これは、クエリと文書の距離に基づいてフィードバックのパラメータを動的に調整することにより、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ことを目指すものである。

ところで、先に述べた、拡張適合フィードバックによるWWW情報検索手法においては、以下のような改善すべき点があった。すなわち、検索結果についての適合／不適合の評価をユーザに要求することからユーザに負荷を与え、検索結果の多くに対してユーザの評価を得ることが容易でなく、その結果ユーザの興味を学習する際に偏りが生じること等である[3, 4]。

このような問題への対処法の一つとして、大量の検索結果に対して文書クラスタリングを行い、インタラクティブに適合情報を絞り込む方法[5, 6, 7]が有効であろう。すなわち、(1) 大量の検索結果をいくつかのグループにクラスタリングし、(2) それに対してユーザが適合と判断した複数のグループを、(3) システムがマージ・再度クラスタリングを行う、といったユーザとのインタラクションを複数回繰り返すことにより、大量の文書からユーザの興味と合致するものを絞り込む手法である。ここで、従来の文書クラスタリングにおいては、文書に固定的に付与された特徴ベクトルのみに基づいて文書の分類が行われていた[8, 9]ため、インタラクションの過程で動的に変化するユーザの興味を反映するものではない。

一方、通常の情報検索の分野において、適合フィードバックを適用する際に修正されたクエリの状態をユーザに提示し、また、ユーザがそれに対して操作可能にすることで、ユーザとの親和性が向上すると共に、検索精度が高まることが指摘されており[10]、適合フィードバックによって修正されたキーワードの候補を、ユーザが適合文書をチェックする度に漸次的に更新するインターフェースが提案されている[11]。しかしながら、フィードバックの度合いが固定された適合フィードバックが用いられているために、ユーザの興味が連続的に変化することが考慮されていない。

本稿では、WWW情報検索における大量の検索結果をクラスタリングし、ユーザの選択した文書クラスターから再クラスタリングすることを繰り返すことにより、ユーザが多く検索結果に対して適合／不適合評価を行うことを支援するが、新たな提案として、クラスタリングの際に文書クラスターに対してユーザが行った評価により漸次的に更新されたクエリを利用して、ユーザの興味に基づいたクラスタリングを行う。ここで、クエリの漸次的な修正は前述の拡張適合フィードバックを用いて、ユーザの興味を反映したものとなっている。なお、クエリの詳細はユーザに提示され、必要に応じてユーザが直接修正することができる。また、ユーザは適宜、以上の

ようにして再構築された新たなクエリにより再検索することもできる。

2 適合フィードバックのユーザへの適応性を考慮した拡張

ユーザの検索態度として、(1) ユーザが明確な検索目標を有する場合、(2) 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興味が絞り込まれていく場合、(3) ユーザが検索過程において新たな知識を発見し検索目標の翻意を行う場合、等が考えられる。このような多様なユーザの検索行為に対処するために、我々は、動的に変化するユーザの検索目標への適応を目指して、伝統的な適合フィードバックを拡張し[2, 3, 4]、WWW情報検索への適用を検討してきた。これは、クエリと文書の距離に基づいてフィードバックのパラメータを動的に自動調整することにより、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ことを図るものである。

3節で述べる本稿で提案する手法において重要な位置を占める要素技術として、以下に、拡張適合フィードバックについて要点を述べる。

2.1 適合フィードバック

本稿ではベクトル空間モデル[12]を用いて、インデクシング[3, 4]により抽出された語集合を基底語とし、各々の基底語に対応する重みを成分とした特徴ベクトルで文書を表現する。

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、クエリのベクトルの重みに寄与させ、検索精度を高めようとするものである。ここでは、Rocchio の式[1]を用いて、これを実現する。

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

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 は検索結果に対する適合文書、不適合であるとわかっている文書の集合を表し、適合とも不適合とも判断できない文書に関しては適合フィードバックの対象としない。 \mathbf{q}_{k+1} において、重みが負の値をとる語は除去する。また、ベクトル $\hat{\mathbf{q}}_k, \hat{\mathbf{d}}_i, \hat{\mathbf{d}}_j$ はそれぞれ正規化されているものとする。

従来、Rocchio の式(1)におけるパラメータ 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[13]。

2.2 適合フィードバックの動的パラメータ調整

ユーザー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) 例えば、ユーザーの翻意が発生した場合では、クエリと適合文書が近接しないことが考えられる。このとき、式(1)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) 例として、ユーザーが明確な検索目標を持ち検索精度の向上を期待している場合では、クエリと適合文書が近接していることが多いと考える。このとき、式(1)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、クエリと不適合文書が近接している場合、式(1)の β は大きくとり、クエリと不適合文書が近接していない場合 β の値は従来用いられてきた 0.5 に近い値をとる。

以上のような考え方のもと、式(1)のパラメータ α, β をクエリ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ろ [3, 4]。

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

$$\beta = \begin{cases} 0.5 & (\max_{\mathbf{d}_j \in N} (\hat{\mathbf{q}}_k, \hat{\mathbf{d}}_j) < b) \\ x_\beta + y_\beta \cdot \max_{\mathbf{d}_j \in N} (\hat{\mathbf{q}}_k, \hat{\mathbf{d}}_j) & (\max_{\mathbf{d}_j \in N} (\hat{\mathbf{q}}_k, \hat{\mathbf{d}}_j) \geq b) \end{cases}. \quad (3)$$

ただし、 $\hat{\mathbf{q}}_k$ および $\hat{\mathbf{d}}_i, \hat{\mathbf{d}}_j$ は正規化されている。また、 $\max_{\mathbf{d}_i \in R} (\hat{\mathbf{q}}_k, \hat{\mathbf{d}}_i) = a$, $\max_{\mathbf{d}_j \in N} (\hat{\mathbf{q}}_k, \hat{\mathbf{d}}_j) = b$ 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。

$(x_\alpha, y_\alpha) = (0.010, 0.722)$, $(x_\beta, y_\beta) = (0.244, 0.756)$, $a = 0.679$, $b = 0.339$ としたときの α, β の特性を図 1 に示す。

3 提案手法の枠組み

提案手法の枠組みを、図 2 に示す。提案手法は、以下の 2 つの要素技術に大別される。ただし、これらは相互に関連している。

- (1) ユーザの興味を反映した対話的な文書クラスタリング： 検索結果を個々の文書間の距離に基づいてクラスタリングする。ここで、ユーザーが適合クラスターを選択し、システムはそれらに基づいて再クラスタリングを行う。このようなインタラクション

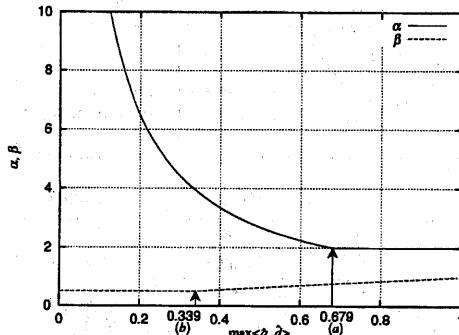


図 1: 適合フィードバックの係数の特性

ンを複数回繰り返すことにより、大量の検索結果からユーザーの興味に適合する文書群を絞り込むことを支援する。3.1.4 項で述べる通り、クラスタリングの際に行う類似度の計算には、ユーザーの興味や観点に基づいた類似性の尺度を新たに導入して、ユーザーの興味を反映した分類を目指す。

- (2) 漸次的なクエリの拡張： クエリはキーワードとその重みの対を要素とした集合として表現されるが、適合フィードバックによって修正されたキーワードの候補を、ユーザーが適合文書（適合クラスター）をチェックする度に漸次的に更新する。このように、多くの検索結果に対するユーザーの評価に基づいて高品質なクエリに修正する。ただし、クエリの詳細はユーザーに提示され、修正されたクエリに対してユーザーがキーワードを取捨選択することを許可する。なお、ユーザーは適宜、修正されたクエリにより、再検索することもできる。

本稿では、新たな提案として、(2) で述べたように文書クラスターに対してユーザーが行った評価により漸次的に更新されたクエリを利用して、(1) においてユーザーの興味に基づいたクラスタリングを行う。また、(1) で行われるユーザーとのインタラクションに応じて、(2) において漸次的にクエリが更新される。ここで、クエリの漸次的な修正には前述の拡張適合フィードバックを用いて、動的に変化するユーザーの興味を反映したものとなっている。

3.1 対話的な文書クラスタリング

適合文書は、不適合文書に対するよりは互いに類似する傾向がある。これはクラスタ仮説(Cluster Hypothesis)[7]と呼ばれる。この仮説に基づいて、検索結果を個々の情報資源間の距離に基づいてクラスタリングする。ここ

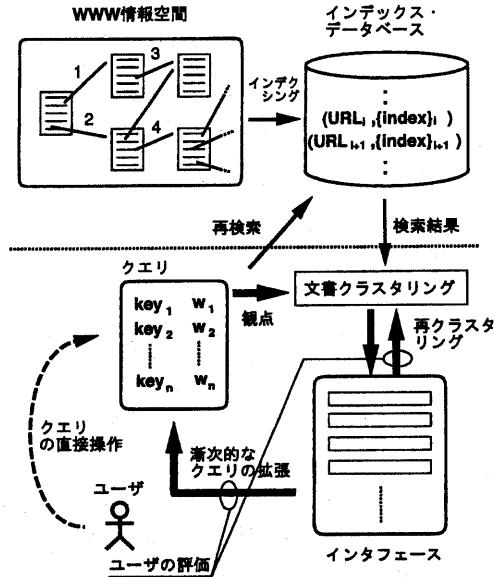


図 2: 提案手法の枠組み

で、ユーザーが適合クラスタを選択し、システムはそれらに基づいて再クラスタリングを行う。このようなインタラクションを複数繰り返すことにより、大量の検索結果からユーザーの興味に適合する情報資源群を絞り込むことができる [5, 6, 7]。

文書クラスタリングには種々の手法が提案されており、文書の集合を階層的に分類する階層的クラスタリング [14] と、特定の分割数へと一気に分割する非階層的クラスタリング [8, 9] に大別される。本稿では、非階層的クラスタリングとして標準的な *k*-means 法を用いる。一般に、*k*-means 法を実現するには、次の 3 つの層を設定する必要がある。

- (1) *k* 個の種子点を発見する。
- (2) 文書のそれぞれを種子点へ配置する。
- (3) 形成された分割を洗練化する。

k 個の分割された文書グループの集合を P とすると、結果は $\bigcup_{\Pi \in P} \Pi = C$ である。

3.1.1 種子点の発見

まず、文書の同一の特徴ベクトルを有する文書が複数存在するとき、一つを除いて残りを削除する。これは、WWWにおいてはミラーサイトが存在すること等の理由で、一つのリンク先について複数の URL が一対多対応していることを考慮している。

本稿では、*k*-means 法における種子点の発見のために、Cutting らにより提案された Fractionation アルゴリズム [8, 9] を用いる。Fractionation では、 n 個の文書を k 個のグループにクラスタリングするための時間計算量は $O(kn)$ であり、インタラクティブな処理のために分類の正確さよりも速さを考慮して設計されている。また、当アルゴリズムは初期種子点を発見することのみを目的としており、実行速度が遅くとも良好にクラスタリングを実現するアルゴリズムの存在を前提としているが、この前提となるアルゴリズムとして、最も類似する 2 つの文書(文書クラスタ²)を一つのクラスタにまとめることを順次繰り返していくってクラスタ数を一つずつ減らしていくという、単純クラスタリングを用いる [15]。ここで用いる類似度の関数には余弦尺度を用い、 d_i, d_j をそれぞれ文書(文書クラスタ)ベクトルとすると、次式で表すことができる。

$$\text{sim}_C(d_i, d_j) = \frac{\langle d_i, d_j \rangle}{\|d_i\| \|d_j\|}. \quad (4)$$

Fractionation アルゴリズムは k 個の種子点を発見するために、最初に n 個の文書集合 C を固定サイズが $m > k$ の、 n/m 個の固定要素数のクラスタに分割する。次に、これらの固定要素数のクラスタ毎に単純クラスタリングが適用され、数の減少率がおよそ ρ となるよう個体をグループ化する。次に、これらのグループを個体とみなして上記の一連の過程を反復する。 k グループだけが残るときに反復が終了する。

3.1.2 文書の種子点への配置

ひとたび k 個の種子点が見つかり、それに適する特徴ベクトルが定義できれば、 C におけるそれぞれの文書は何らかの評価基準に基づいてそれらの種子点のいずれかに配置される。この評価基準で最も単純なアルゴリズムは、それぞれの文書を最も近い種子点に配置するものであるが、本稿ではこれを採用する。

G を k 個のグループへの分類の集合であるとする。また、 Γ_i は G における i 番目のグループの重心ベクトルであるとする。文書 d と Γ_i について、余弦尺度、

$$\text{sim}_C(d, \Gamma_i) = \frac{\langle d, \Gamma_i \rangle}{\|d\| \|\Gamma_i\|}. \quad (5)$$

が最大であるなら、 $d \in \Pi_i$ とする。そのとき、集合 $P = \{\Pi_i\}, 0 \leq i \leq k$ は所望の分割となる。この手続きの時間計算量は $O(kn)$ である。

3.1.3 洗練化

初期クラスタリングが与えられたら、次にそれをより精度の高いクラスタリングに洗練化することが必要である。

²文書クラスタの特徴ベクトルとして、そのクラスタに属する複数の文書ベクトルの重心をとる。

る。上述の初期クラスタリングアルゴリズムに同様に、ここでも速さと正確さのトレードオフが存在する。本稿では、最も単純な手順であるが、種子点への配置と同様にそれぞれの文書を最も近いクラスタに最適配置することにより実現する。

与えられたクラスタの集合から、クラスタの重心によってクラスタの種子点を生成する。次に、新しい文書クラスタを形成するために、それぞれの文書を最も近い種子点に配置する。この手順は盲目的に繰り返されるため、少い固定回数のみ繰り返される。

3.1.4 ユーザの興味を反映した文書クラスタリング

3.1.1項から3.1.3項までに述べてきた文書クラスタリングは、文書に固定的に付与された特徴ベクトルをもとにに行われてきた[8, 9]が、動的に変化するユーザの興味に基づいてクラスタリングが実現されることが望ましい。

さて、視点や観点による類似性の違いに着目した研究がいくつかある[16, 17]が、いずれもシソーラスなどの知識体系を用いているため、大量の情報を扱う情報検索には適さない。本稿では、クエリを構成するキーワードと類似度の計算の対象である文書ベクトル(あるいは文書クラスタの重心ベクトル)を成すキーワードとの表層的なマッチングを行い、文書ベクトルで0でない成分を持つキーワードの成分に、そのキーワードについてのクエリベクトルにおける成分に比例する値を加算することにより文書ベクトルを変調し、それを用いて類似度を求める。すなわち、

$$\text{sim}_V(\mathbf{d}_i, \mathbf{d}_j) = \frac{\langle \tilde{\mathbf{d}}_i, \tilde{\mathbf{d}}_j \rangle}{\|\tilde{\mathbf{d}}_i\| \|\tilde{\mathbf{d}}_j\|}. \quad (6)$$

ここで、ベクトル $\tilde{\mathbf{d}}_i, \tilde{\mathbf{d}}_j$ の基底語 t に関する成分 $w_i^{d_i}, w_j^{d_j}$ はそれぞれ次式により求める。

$$\tilde{w}_i^{d_i} = \begin{cases} w_i^{d_i} + \xi w_i^q & (w_i^{d_i} \neq 0) \\ w_i^{d_i} & (w_i^{d_i} = 0) \end{cases}, \quad (7)$$

$$\tilde{w}_j^{d_j} = \begin{cases} w_j^{d_j} + \xi w_j^q & (w_j^{d_j} \neq 0) \\ w_j^{d_j} & (w_j^{d_j} = 0) \end{cases}. \quad (8)$$

但し、 $w_i^{d_i}, w_j^{d_j}$ はそれぞれ文書ベクトル $\mathbf{d}_i, \mathbf{d}_j$ を正規化したときの基底語 t に関する成分であり、 w_i^q はクエリベクトル \mathbf{q} を正規化したときの基底語 t に関する成分である。また、 ξ は適当な係数であり、本稿では1を用いた。なお、クエリはユーザが文書クラスタについて適合／不適合をチェックする度に、選択された文書クラスタの重心ベクトルをクエリに加えることで更新される³。修正されたクエリはユーザの興味を反映していると期待される。このため、3.1.1項から3.1.3項までに述べた文

³ただし、クエリをユーザが直接操作することも許可されている。

書クラスタリングにおける類似度計算において、式(4), (5)で示される類似度関数 sim_C の代わりに式(6)を用いることによって、ユーザの興味や観点を考慮した文書クラスタリングが実現できると考える。

3.1.5 クラスタの要約生成

クラスタの重心ベクトルを構成するキーワード群の内、重みの順に上位のものをクラスタキーワードとする。その他に、クラスタに含まれる文書の内、クラスタの重心と最も類似しているものを発見し、そのタイトルを典型タイトルとする。これらをクラスタの要約として、ユーザに提示する[8]。

ここで、本稿における新たな提案として、各々の文書とクラスタの重心との類似度の計算において、式(6)で、 \mathbf{d}_i をクラスタ内の文書ベクトル、 \mathbf{d}_j をクラスタの重心ベクトルとすることにより、最も類似度の高い文書を選び出す。これにより、クラスタの特徴を示す代表的な文書であり、かつ、ユーザの興味にも比較的近い文書のタイトルをクラスタの要約とことができ、ユーザの求める情報がクラスタ内に含まれているかどうかを判定することを促すことができるものと考える。

3.2 漸次的なクエリの拡張

クエリの拡張には、シソーラスによる概念的類似性の高いキーワードを用いる方法、過去のクエリの履歴を統計的に処理し相関性の高いキーワードを提示する方法[18]、適合フィードバックにより検索結果においてユーザが適合評価した文書の特徴情報をクエリにフィードバックする方法などがある。適合フィードバックを適用する際に修正されたクエリの状態をユーザに提示し、また、ユーザがそれに対して操作可能にすることで、ユーザとの親和性が向上すると共に、検索精度が高まることが指摘されており[10]、適合フィードバックによって修正されたキーワードの候補を、ユーザが適合文書をチェックする度に漸次的に更新するインターフェースが提案されている[11]。しかしながら、フィードバックの度合いが固定された適合フィードバックが用いられているために、ユーザの興味が連続的に変化することが考慮されていない。

本稿では、連続的に変化するユーザの検索目標に柔軟に適応することを目指した拡張適合フィードバックを用い、更に、修正されたクエリの状況をユーザに提示し、ユーザによる修正を可能にすることで、ユーザに対する親和性を向上させることを目指す。

また、従来の手法は、単に個々の適合文書から得られる情報からクエリを修正していたが、本稿では、文書クラスタリングを併用するため、適合クラスタの特徴情報をもとにクエリを修正する。ここで、対話的な文書クラ

スタリングにおけるインタラクションから自動的にクエリが修正されるため、ユーザが大量の検索結果から適合／不適合の評価を行うことを支援することができる。

なお、修正されたクエリは時間的に変化するユーザの興味を表現したものとなり、3.1.4および3.1.5項に述べるように文書クラスタリングやその要約生成にも利用され、また、それを新たなクエリとして、適宜再検索することもでき、検索精度の向上が期待される。

4 システムの実装および実験

4.1 対象領域

本稿では、ロボット[19]と呼ばれるソフトウェアモジュールを用いて、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。

4.2 インタフェース

インターフェースにはJavaを用いて計算機上に実装した。これにより、検索結果に対するインタラクションの処理をブラウザ側で実行することができ、サーバの負荷を軽減することができる。プロトタイプシステムのインターフェースを図4.2に示す。

ユーザは、まずクエリをシステムに入力する。これに対して、システムはクエリとインデックス・データベースに格納された文書とを照合して、クエリ・文書間の類似度[3, 4]による順位付けにおいて上位の文書群を対象に、文書数が閾値以上であれば3.1節で述べた文書クラスタリングが行われ、その分類結果を図4.2(a)に示すように表示し、閾値以下であれば図4.2(b)のように文書群をフラットに表示する。本稿ではこの閾値を20とした。

図4.2(a)に示すように、文書クラスタの表示においては、内包する文書数とそれを表示するためのボタン「View」、クラスタの要約としての典型的な文書のタイトルとキーワード、ユーザ評価用のボタンを1セットとしてユーザに提示する。ユーザはクラスタの要約やクラスタが内包する文書を閲覧することによって、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。「Re-Clustering」のボタンをマウスクリックすることで適合評価された複数の文書クラスタに含まれる文書群を対象に、文書数が閾値以上であれば再クラスタリングを行い図4.2(a)と同様な分類表示を、

閾値以下であれば図4.2(b)に示すようなフラットな表示を行う。

図4.2(b)に示すように、フラットな表示においては、類似度の順位、類似度、タイトル、キーワード、URL、ファイルサイズ、最終更新日、およびユーザ評価用のボタンを1セットとして類似度順に表示する。ここで、URLは実際のWWWページへとリンクされている。ユーザは、図4.2(a)と同様、適合、不適合、わからないの評価に応じて、「Good」、「NG」、「?」のいずれかのボタンを選択する。

なお、図4.2(a),(b)のいずれの表示においても、「Query-Window Open」、「Query-Window Close」のボタンをマウスクリックすることで、キーワードとその重要度からなるクエリ内容が記述された別枠のウインドウが開閉する。ただし、このウインドウ内の左のカラムは検索時に用いたクエリ、右のカラムは検索時に用いたクエリを初期値とするが、ユーザが文書／文書クラスタの適合／不適合を評価する度に、3.2節で述べたように漸次的に修正され再表示される。クエリを構成するキーワードをユーザが選択し、「Delete」ボタンをマウスクリックすることで削除することもできる。なお、右カラムのクエリの上方には2.2節で述べた拡張適合フィードバックによって自動調整される適合フィードバックのパラメータが表示されている。また、「Re-Search」のボタンをマウスクリックすることでユーザの評価により修正されたクエリを用いて再検索が行われる。

4.3 実験

実験のため、まず、4.1節で述べたインデックス・データベースで適合文書数の比較的多い「Neural Network」をクエリとして、拡張適合フィードバックによる再検索を2回繰り返したときの上位200件を検索結果とする。検索結果の文書群を対象に、従来の式(4),(5)で示される類似度関数 sim_C を用いて文書クラスタリングを行つた。このときの典型タイトルやクラスタキーワードを表1に示す⁴。ただし、典型タイトルの括弧内の数字は、そのクラスタに内包される文書数を示している。次に、文書クラスタ9を適合とみなし、再クラスタリングを行う。このとき、先ほどと同様、 sim_C を用いる場合と、式(6)で示される類似度関数 sim_V を用いる場合のそれについて分類結果を比較する。それぞれの分類結果を表2に示す。この実験では、「Neural Network」に関する検索結果の内、文書クラスタ9のキーワードからわかる通り、「学習」について絞り込みを行うことになる。このときのユーザの興味は表3で示されるクエリにより表現され、 sim_V を用いる場合はこのクエリを類似度に寄与されることになる。表3のクエリを成すキーワードか

⁴表1では特定の個人名、団体名は可能な限り匿名とした。表2および表3も同様である。

表 1: sim_C を用いた分類例の抜粋

	典型タイトル	クラスタキーワード
...		
6	(5) Laboratories of Depts. of EEICE	羽島, 相澤, 近山, 電気, 鳥, 山地, 系, ...
7	(10) 前川研 動物園	前川, maekawa, ファジー, 動物園, グループ, ...
8	(5) Information on Computers in Depts. of EE	電気, depts, electrical, professors, courses, 総合, ...
9	(25) K'S ROOM	学習, 異種, kawaishi, fai, mild, 一貫, csp, 認識, ...
10	(15) Related URL	linux, conference, 機業, intelligence, artificial, ...
...		

表 2: sim_C , sim_V を用いた分類結果の比較

sim_C を用いる場合				
...	クラスタ 1	クラスタ 2	クラスタ 3	...
	K 氏	T 研究室	Ni 研究室	
	K 氏	Y 研究室	Ni 研究室	
	T 研究室	Ta 大学	-	
	R 大学	-	-	
	Nu 研究室	-	-	

sim_V を用いる場合				
...	クラスタ 1	クラスタ 2	クラスタ 3	...
	T 研究室	Ni 研究室	Y 研究室	
	T 研究室	Ni 研究室	Y 研究室	
	T 研究室	-	Ta 大学	
	K 氏	-	Ts 大学	
	K 氏	-	R 大学	
	-	-	Nu 研究室	

らわかるように、「学習」以外のキーワードは各々の研究室や研究者、研究テーマを表現するものとなっている。表 2 の分類結果より、 sim_C を用いた従来手法による文書クラスタリングでは、T 研究室や Y 研究室のそれについて、類似する⁵にも関わらず異なる文書クラスタに分類されており、 sim_V を用いた提案手法ではそれらが同一のクラスタに分類されている。このことから、3.1.4 項で述べたユーザの興味を反映した文書クラスタリングの有意性が確認される。

5 おわりに

本稿では、新たな対話的文書クラスタリング手法として、大量の検索結果のクラスタリングの際に文書クラスタに対してユーザが行った評価により漸次的に更新されたクエリを利用して、ユーザの興味に基づいたクラスタ

⁵ 第一著者が実際のページを閲覧して、本実験に関しては同一の研究室のページは類似していることを確認した。

表 3: 実験において用いられたクエリの詳細

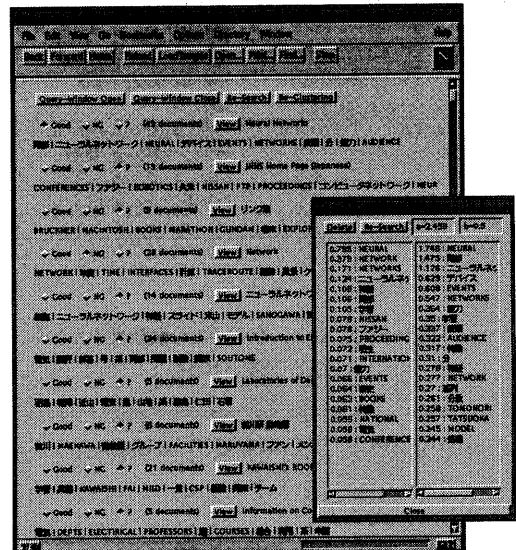
キーワード	重み	キーワード	重み
学習	10.263	一貫	1.892
異種	2.801	csp	1.845
K 氏	2.778	認識	1.761
fai	2.359	探索	1.730
mild	2.114	...	

リングを行い、ユーザの選択した文書クラスタから再クラスタリングすることを繰り返す手法を提案した。これにより、大量の検索結果に対して適合情報を絞り込むことができ、また、それをもとに修正した高品質なクエリにより、再検索することにより検索精度の向上が期待できる。ここで、拡張された適合フィードバックを用いて、動的に変化するユーザの興味を反映したクエリの漸次的な修正を実現している。また、クエリの状態はユーザに提示され、必要に応じてユーザが修正ができる。本稿では基礎的な実験結果により提案手法の有効性を確認したが、詳細な検討は今後の課題である。

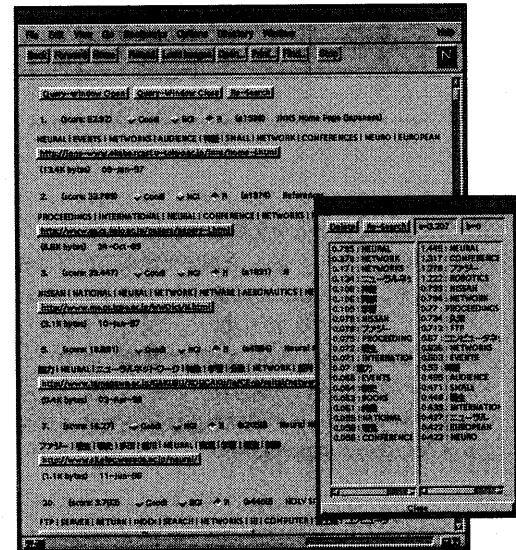
参考文献

- [1] Rocchio, J. J.: Relevance Feedback in Information Retrieval, *The SMART Retrieval System : Experiments in Automatic Document Processing* (Salton, G.(ed.)), Prentice Hall, pp. 313–323 (1971).
- [2] 江口浩二, 藤本剛司, 山口洋介, 伊藤秀隆, 横元昭: 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 年電子情報通信学会総合大会, No. D-4-1, p. 74 (1997).
- [3] 江口浩二, 藤本剛司, 伊藤秀隆, 横元昭: ユーザへの適応性を考慮した WWW 情報検索, 電子情報通信学会 第 8 回データ工学ワークショップ (DEWS'97), pp. 203–208 (1997).
- [4] Eguchi, K., Ito, H. and Kumamoto, A.: Information Retrieval Considering Adaptation to User's Behaviors on the WWW, *The 2nd International Workshop on Information Retrieval with Asian Languages (IRAL'97)* (to appear).
- [5] 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).
- [6] Pirolli, P., Schank, P., Hearst, M. A. and Diehl, C.: Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection, *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'96)*, pp. 213–220 (1996).

- [7] Hearst, M. A. and Pedersen, J. O.: Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, *Proc. ACM SIGIR '96*, pp. 76–84 (1996).
- [8] Cutting, D. R., Karger, D., Pedersen, J. O. and Tukey, J. W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, *Proc. ACM SIGIR '92*, pp. 318–329 (1992).
- [9] Cutting, D. R., Karger, D. and Pedersen, J. O.: Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections, *Proc. ACM SIGIR '93*, pp. 126–134 (1993).
- [10] Koenemann, J. and Belkin, N. J.: A case for interaction: A study of interactive information retrieval behavior and effectiveness, *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'96)*, pp. 205–212 (1996).
- [11] Beaulieu, M.: Experiments on interfaces to support query expansion, *Journal of Documentation*, Vol. 53, No. 1, pp. 8–19 (1997).
- [12] Salton, G.: *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley (1989).
- [13] 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)* (Harman, D. K.(ed.)), NIST Special Publication 500-215, pp. 45–56 (1994).
- [14] Willett, P.: Recent trends in hierachic document clustering: A critical review, *Information Processing & Management*, Vol. 24, No. 5, pp. 577–597 (1988).
- [15] 長尾真: パターン情報処理, コロナ社 (1983).
- [16] 沢田裕司, 大川剛直, 馬場口登: 観点を考慮した連想機構の実現, 情報処理学会論文誌, Vol. 35, No. 5, pp. 714–724 (1994).
- [17] 笠原要, 松澤和光, 石川勉, 河岡司: 観点に基づく概念間の類似性判別, 情報処理学会論文誌, Vol. 35, No. 3, pp. 505–509 (1994).
- [18] 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).
- [19] Koster, M.: World Wide Web Robots, Wanderers, and Spiders, <http://info.webcrawler.com/mak/projects/robots.html> (1996).



(a) 文書クラスタの表示



(b) フラットな文書表示

図 3: プロトタイプシステムのインターフェース