データベースシステム 114－7
（1998. 1. 19）

検索結果のクラスタリングに基づくユーザへの適応性を考慮した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 本稿では，動的に変化するユーザの興味を反映することを考慮した，対話的文書クラスタリング手法に基づく適合フィードバックの新たな拡張手法を提案する．大量の検索結果に対してクラスタリングが行われ，文書クラスタに対してユーザが適合評価を行うことにより，ユーザの適合評価に要求される負荷が軽減される．提案手法により漸次的に修正された高品質なクエリを用いて再検索を行なうことにより，検索効率の向上が期待できる．

Incremental Query Expansion Considering Adaptation to User's Behavior Based on Clustering the Search Results

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 This paper proposes a new extension of the relevence feedback, based on interactive document clustering, for reflecting dynamically changing interests of users of information retrieval systems. In the proposed method, the users evaluate the relevance of document clusters, instead of individual documents, which reduces the load of the users. The incrementally expanded and refined queries are used in re-searching to improve the retrieval effectiveness.

1 はじめに

近年，インターネットの普及にともない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文書情報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(以下，拡張適合フィードバック) を提案し，検討してきた[2, 3, 4]．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，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，一般に，適合フィードバックによる情報検索手法においては，以下のような問題点がある．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．

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] と組み合わせることが有効であると考える．これは，(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．

一方，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[8]，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フェースが提案されている[9]．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．また，ユーザの評価は適合文書単体に対して行われることを前提としているが，クラスタリングされた検索結果に対する評価からユーザの興味を学習するためには，新たな拡張を行う必要がある．

本稿では，大量の検索結果に対してクラスタリングを行ない，そこでユーザが適合と評価したクラスタからユーザの興味を的確に学習することを目指して，適合フィードバックを拡張し，漸次的にクエリを修正する手法を提案する[10]．このとき，ユーザの検索目標が漠然としていたり，動的に変化する場合などを考慮する．本稿では，提案手法をWWW情報検索に適用し，そのプロトタイプシステムについての説明を加える．

[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．

2 提案手法の枠組み

提案手法の枠組みを，図1に示す．ここでは，次の2つの要素技術を用いる．

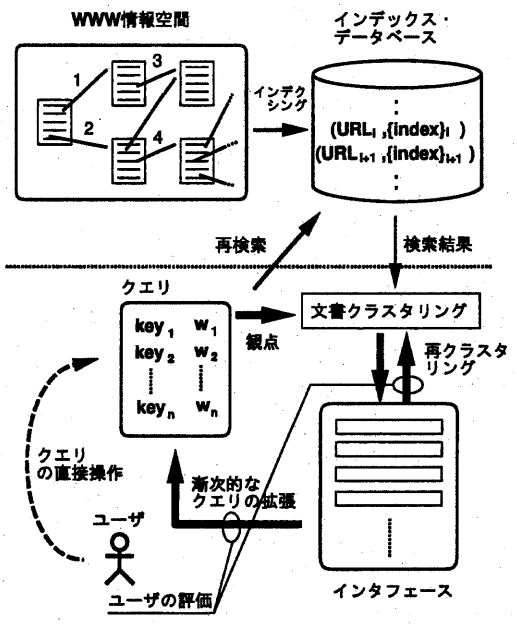

図 1: 提案手法の枠組み

(1) 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，ユーザが適合クラスタを選択し，システムはそれらに含まれる文書群に対し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回繰り返すことにより，大量の検索結果からユーザの興味に適合する文書群を絞り込むことを支援する．

(2) 適合フィードバックによる漸次的なクエリの修正　本稿では，クエリを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するが，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，ユーザが適合文書クラスタ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に提示され，修正されたクエリに対してユーザがキーワードを取捨選択することを許可する．このようにして修正されたクエリを用いて，ユーザは適宜，サーチエンジンサーバから再検索することができる．

提案手法では，(1) で行われるユーザ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おり，クエリは動的に変化するユーザの興味を反映したものとなっている[2]．

2.1節，2.2節として，上記の個々の要素技術について説明を加える．

2.1 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，検索結果を個々の文書ベクトル間の距離に基づいてクラスタリングする．ここで，ユーザ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]．

クラスタリング・アルゴリズムには種々のものが提案されているが，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．

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 [12, 13] 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 つの文書または文書クラスタ[3]を一つのクラスタにまとめることを順次繰り返してい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 [14] を用いる．

[2]文書クラスタに対してユーザが行った適合評価により漸次的に更新されたクエリの情報は，そのときのユーザの興味を反映したものとなっているものと思われる．著者は，これを積極的に活用し，(1) におけるクラスタリングの際に行う類似度の計算には，クエリの情報をユーザの興味や観点とみなし，これに基づいた類似性の尺度を新たに提案して，ユーザの興味を反映したクラスタリングの実現を目指している [10, 11]．これに関しては，本稿では割愛する．

[3]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．

2.2 適合文書または不適合文書からのユーザの興味の学習

2.2.1 Rocchio の式

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．適合フィードバックを実現する手法は種々あるが，以下に Rocchio の式 [1] を示す．

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

ただし，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，α, β はフィードバックの度合いを示すパラメータである．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．

2.2.2 ユーザへの適応性を考慮した適合フィードバック

著者はこれまで，ユーザの検索目標が漠然としている場合や動的に変化する場合に対処するため，式(1) に示された適合フィードバックのパラメータ α, β を従来のようにそれぞれ 2, 0.5 といった値に固定するのでなく動的に調整する拡張手法を検討してきた [2, 3, 4]．そこでは，次の仮定を設けている．まず，適合評価について，

(1)例えば，ユーザ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る．

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

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

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，$\max_{\hat{\mathbf{d}}_i \in R} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{d}}_i \rangle = a$, $\max_{\hat{\mathbf{d}}_j \in N} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{d}}_j \rangle = b$ 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．

式(1)のパラメータ α, β をそれぞれ2.0, 0.5に固定した場合と，式(2),(3)において $(x_\alpha, y_\alpha) = (0.010, 0.722)$, $(x_\beta, y_\beta) = (0.244, 0.756)$ としてパラメータ α, β を調整した場合とを，ユーザの検索目標が動的に変化することを想定した実験により比較したところ，平均適合率において 29.7% の精度向上が見られた [3].

3 クラスタリングに基づく適合フィードバック

本章では，今回新たに提案する，対話的クラスタリングに基づく，適合フィードバックの拡張手法について述べる．

3.1 クラスタリングに基づく Rocchio の式の拡張

2.2.1項で述べた従来の適合フィードバック手法では，ユーザの評価は適合文書単体に対して行われることを前提としているが，2.1節で述べたような文書クラスタリングにおける適合クラスタに対する評価からユーザの興味を学習するためには，適合フィードバックに拡張を行う必要がある．(1) 式を次のように拡張する．

$$\mathbf{q}_{k+1} = \hat{\mathbf{q}}_k + \frac{\alpha}{|RC|} \sum_{G_r \in RC} \hat{\mathbf{c}} - \frac{\beta}{|NC|} \sum_{G_n \in NC} \hat{\mathbf{c}}, \quad (4)$$

$$\hat{\mathbf{c}}(G_r) = \frac{\mathbf{c}(G_r)}{\|\mathbf{c}(G_r)\|}, \ \hat{\mathbf{c}}(G_n) = \frac{\mathbf{c}(G_n)}{\|\mathbf{c}(G_n)\|}, \quad (5)$$

$$\mathbf{c}(G_r) = \sum_{\hat{\mathbf{d}}_i \in G_r} \hat{\mathbf{d}}_i, \quad (6)$$

$$\mathbf{c}(G_n) = \sum_{\hat{\mathbf{d}}_j \in G_n} \hat{\mathbf{d}}_j. \quad (7)$$

ここで，RC, NC はそれぞれクラスタリングされた検索結果に対してユーザにより適合評価されたクラスタ G_r，不適合評価されたクラスタ G_n の集合を表す．

(4) 式は，著者がこれまで用いてきたものである [10] が，少数の文書からなるクラスタと多数の文書からなるクラスタについて，それぞれの学習情報に同等の比率を与えているため適切でないものと思われる．そこで，本稿では新たに次式を提案する．

$$\mathbf{q}_{k+1} = \hat{\mathbf{q}}_k + \frac{\alpha}{|\cup_{G_r \in RC} G_r|} \sum_{G_r \in RC} \sum_{\hat{\mathbf{d}}_i \in G_r} \hat{\mathbf{d}}_i - \frac{\beta}{|\cup_{G_n \in NC} G_n|} \sum_{G_n \in NC} \sum_{\hat{\mathbf{d}}_j \in G_n} \hat{\mathbf{d}}_j. \quad (8)$$

3.2 クラスタリングに基づいた適合フィードバックの動的パラメータ調整

2.2.2項にて，ユーザへの適応性を考慮して動的に適合フィードバックのパラメータを調整する手法について述べた．本稿では，このような適合フィードバックの動的パラメータ調整法を，2.1節で述べたような文書クラスタリングに基づいて拡張する．以下に提案する関数は式(8)のパラメータ α, β を与えるものである．

3.2.1 関数 A

まず考えられることが，次式のように，クラスタの重心ベクトルを用いることである．

$$\alpha = \begin{cases} 1/\left(x_\alpha + y_\alpha \cdot \max_{G_r \in RC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{c}}(G_r) \rangle\right) \\ \qquad (\max_{G_r \in RC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{c}}(G_r) \rangle \leq a) \\ 2 \qquad (\max_{G_r \in RC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{c}}(G_r) \rangle > a) \end{cases}, \quad (9)$$

$$\beta = \begin{cases} x_\beta + y_\beta \cdot \max_{G_n \in NC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{c}}(G_n) \rangle \\ \qquad (\max_{G_n \in NC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{c}}(G_n) \rangle \geq b) \\ 0.5 \qquad (\max_{G_n \in NC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{c}}(G_n) \rangle < b) \end{cases}, \quad (10)$$

$$\hat{\mathbf{c}}(G_r) = \frac{\mathbf{c}(G_r)}{\|\mathbf{c}(G_r)\|}, \ \hat{\mathbf{c}}(G_n) = \frac{\mathbf{c}(G_n)}{\|\mathbf{c}(G_n)\|}, \quad (11)$$

$$\mathbf{c}(G_r) = \sum_{\hat{\mathbf{d}}_i \in G_r} \hat{\mathbf{d}}_i, \quad (12)$$

$$\mathbf{c}(G_n) = \sum_{\hat{\mathbf{d}}_j \in G_n} \hat{\mathbf{d}}_j. \quad (13)$$

上式は，著者がこれまで用いてきたものである [10] が，比較的多数の文書からなるクラスタにおいては，不適合文書をも含む可能性が高く，このとき，クラスタの重心ベクトルはクエリと類似しない傾向にあり，α の値は大きくなる傾向に，β の値は小さくなる傾向にあると考えられる．

3.2.2 関数 B

関数 A の問題に対処するために，クラスタの重心ベクトルの代わりに，クラスタ内でクエリと近いいくつかの文書の重心ベクトルを用いることが考えられる．

$$\alpha = \begin{cases} 1/\left(x_\alpha + y_\alpha \cdot \max_{G_r \in RC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(G_r) \rangle\right) \\ \qquad (\max_{G_r \in RC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(G_r) \rangle \leq a) \\ 2 \qquad (\max_{G_r \in RC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(G_r) \rangle > a) \end{cases}, \quad (14)$$

$$\beta = \begin{cases} x_\beta + y_\beta \cdot \max_{G_n \in NC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(G_n) \rangle \\ \qquad (\max_{G_n \in NC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(G_n)_j \rangle \geq b) \\ 0.5 \qquad (\max_{G_n \in NC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(G_n)_j \rangle < b) \end{cases}, \quad (15)$$

$$\hat{\mathbf{s}}(G_r) = \frac{\mathbf{s}(G_r)}{\|\mathbf{s}(G_r)\|}, \ \hat{\mathbf{s}}(G_n) = \frac{\mathbf{s}(G_n)}{\|\mathbf{s}(G_n)\|}, \quad (16)$$

$$\mathbf{s}(G_r) = \sum_{\hat{\mathbf{d}}_i \in \Lambda^m_{G_r}} \hat{\mathbf{d}}_i, \quad (17)$$

$$\mathbf{s}(G_n) = \sum_{\hat{\mathbf{d}}_j \in \Lambda^m_{G_n}} \hat{\mathbf{d}}_j. \quad (18)$$

ここで，ある文書集合 G に対してその部分集合Λ_G^m ($\Lambda_G^m \subseteq G$) は，$\langle \hat{\mathbf{q}}_k, \hat{\mathbf{d}}_i \rangle$ ($\hat{\mathbf{d}}_i \in G$) の値が大きいものから順に m 個の文書ベクトルからなる集合とする．

式(17),(18)において $m=1$ のとき，式(8),(14),(15)の関数は式(1),(2),(3)で表される式において $R = \cup_{G_r \in RC}$, $N = \cup_{G_n \in NC}$ としたときの関数と等価である．

なお，式(14),(15)は，適合フィードバックによる学習の度にクラスタ内を走査して，$\mathbf{s}_i$, $\mathbf{s}_j$ を計算する必要があるため，式(9),(10)と比較して計算に手間がかかる．

3.2.3 関数 C

選択された適合クラスタをマージすることによって生成される文書群に対する，クエリと類似する幾つかの文書の重心ベクトルを用いること考えられる．

$$\alpha = \begin{cases} 1/\left(x_\alpha + y_\alpha \cdot \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(\cup_{G_r \in RC}) \rangle\right) \\ \qquad (\langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(\cup_{G_r \in RC}) \rangle \leq a) \\ 2 \qquad (\langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(\cup_{G_r \in RC}) \rangle > a) \end{cases}, \quad (19)$$

$$\beta = \begin{cases} x_\beta + y_\beta \cdot \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(\cup_{G_n \in NC}) \rangle \\ \qquad (\langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(\cup_{G_n \in NC}) \rangle \geq b) \\ 0.5 \qquad (\langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(\cup_{G_n \in NC}) \rangle < b) \end{cases}, \quad (20)$$

$$\hat{\mathbf{s}}(\cup_{G_r \in RC}) = \frac{\mathbf{s}(\cup_{G_r \in RC})}{\|\mathbf{s}(\cup_{G_r \in RC})\|}, \quad (21)$$

$$\hat{\mathbf{s}}(\cup_{G_n \in NC}) = \frac{\mathbf{s}(\cup_{G_n \in NC})}{\|\mathbf{s}(\cup_{G_n \in NC})\|}, \quad (22)$$

$$\mathbf{s}(\cup_{G_r \in RC}) = \sum_{\hat{\mathbf{d}}_i \in \Lambda^m_{(\cup_{G_r \in RC})}} \hat{\mathbf{d}}_i, \quad (23)$$

$$\mathbf{s}(\cup_{G_n \in NC}) = \sum_{\hat{\mathbf{d}}_j \in \Lambda^m_{(\cup_{G_n \in NC})}} \hat{\mathbf{d}}_j. \quad (24)$$

関数 B と同様，式(21),(22)において $m=1$ のとき，式(19),(20)による関数は式(1),(2),(3)で表される関数と等価である．

4 プロトタイプシステム

提案手法を WWW 情報検索に適用し，そのプロトタイプには Java を用いて計算機上に実装した．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文書クラスタリングが行われ，その分類結果を図2に示すように表示し，閾値以下であれば通常のサーチエンジンと同様に文書群をフラットに表示する．

図2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

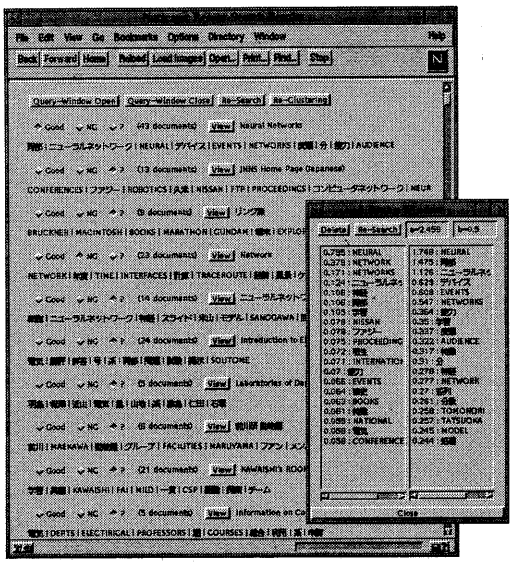

図 2: プロトタイプシステムのインタフェース

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，閾値以下であればフラットに表示する．

なお，図2において，「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ィンドウが開閉する．ただし，このウィンドウ内の左のカラムは検索時に用いたクエリ，右のカラムで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，即時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3.2節で述べた提案手法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を行うことができる．

ロボット [15] 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，プロトタイプシステムにより実験を行ったところ，対話的クラスタリングによるインタラクションと同時にクエリが修正され，これを用いて再検索する

ことによりより多くの適合文書が得られることを確認した．今後，詳細な実験に基づく客観的な評価を行っていく．

5 おわりに

本稿では，動的に変化するユーザの興味を反映することを考慮した，対話的文書クラスタリング手法に基づく適合フィードバックの新たな拡張手法を提案した．大量の検索結果に対してクラスタリングが行われ，文書クラスタに対してユーザが適合評価を行うことにより，ユーザの適合評価に要求される負荷が軽減される．漸次的に修正された高品質なクエリを用いて再検索を行なうことにより，検索効率の向上が期待できる．

今後の課題を以下に列挙する．

(1)本稿で提案したクラスタリングに基づく適合フィードバックの拡張手法について，詳細な実験および客観的な評価を行い，関数の最適化を行う必要がある．

(2)文書クラスタに対してユーザが行った適合評価により漸次的に更新されたクエリの情報は，そのときのユーザの興味を反映したものとなっているものと思われるが，対話的なクラスタリングの際に行う類似度の計算にこれを活用することによる，ユーザの興味を反映したクラスタリング手法を提案している [10, 11]．今後，本稿で提案した手法を用いて，これを有効に機能させることを考えている．

(3)文書クラスタの内容をユーザに提示する方法についても検討していく．

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, 伊藤秀隆, 限元昭: 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) 論文集, pp. 203–208 (1997).

[3]Eguchi, K., Ito, H. and Kumamoto, A.: Information Retrieval Considering Adaptation to User's Behaviors on the WWW, *Proceedings of the 2nd International Workshop on Information Retrieval with Asian Languages (IRAL'97)*, pp. 108–113 (1997).

[4]江口浩二, 伊藤秀隆, 限元昭: 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, 電気学会論文誌 C, Vol. 117-C, No. 11, pp. 1643–1649 (1997).

[5]Hearst, M. A., Karger, D. and Pedersen, J. O.: Scatter/Gather as a Tool for Navigation of Retrieval Results, *Working Notes of the 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, *Proceedings of the 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, *Proceedings of the 19th Annual International ACM SIGIR Conference*, pp. 76–84 (1996).

[8]Koenemann, J. and Belkin, N. J.: A case for interaction: A study of interactive information retrieval behavior and effectiveness, *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'96)*, pp. 205–212 (1996).

[9]Beaulieu, M.: Experiments on interfaces to support query expansion, *Journal of Documentation*, Vol. 53, No. 1, pp. 8–19 (1997).

[10]江口浩二, 伊藤秀隆, 限元昭: 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, 情報処理学会研究報告, FI47-11/NL121-19, pp. 135–142 (1997).

[11]江口浩二, 伊藤秀隆, 限元昭: WWW におけるユーザへの適応性を考慮した対話的文書クラスタリング, 平成 9 年電気関係学会関西支部連合大会講演論文集, No. G15-4, p. G337 (1997).

[12]Cutting, D. R., Karger, D., Pedersen, J. O. and Tukey, J. W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections, *Proceedings of the 15th Annual International ACM SIGIR Conference*, pp. 318–329 (1992).

[13]Cutting, D. R., Karger, D. and Pedersen, J. O.: Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections, *Proceedings of the 16th Annual International ACM SIGIR Conference*, pp. 126–134 (1993).

[14]長尾真: パターン情報処理, コロナ社 (1983).

[15]Koster, M.: World Wide Web Robots, Wanderers, and Spiders, http://info.webcrawler.com/mak/projects/robots/robots.html (1996).