クラスタに基づくブラウジングにおける適応的かつ漸次的なクエリの拡張

1U－9

江口 浩二　伊藤 秀隆　隈元 昭　金田 彌吉
関西大学　工学部

1 はじめに

情報検索における検索結果過多の問題や適合フィードバックにおいて多くの適合性評価が要求される問題等に対処するための新たな提案として，適応パラメータを有する適合フィードバック（adaptive relevance feedback: ARF）[1] とクラスタに基づくブラウジング（cluster-based browsing: CBB）[2] を要素技術とした，文書クラスタリングに基づく適応的かつ漸次的なクエリの拡張（adaptive and incremental query expansion based on document clustering: AIQEC）を提案する [3]．CBB とは，(1) システムが検索結果をクラスタリングし，(2) ユーザが適合であると評価したクラスタ（以下，適合クラスタと呼ぶ）に含まれる文書群に対して，(3) システムが再クラスタリングを行うといった，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適合情報を得ることを支援する手法である．

一般に，従来の適合フィードバックは個々の文書に対する適合性評価に基づいていたが，本稿で提案する AIQEC はクラスタリングされた検索結果に対するユーザの適合性評価からユーザの興味を正確に学習することを目指す．これにより適合性評価に要するユーザの負荷が軽減される．CBB の過程において漸次的に拡張・洗練化されたクエリは随時，検索精度の改善を目的とした再検索に利用され得る．更に，フィードバックパラメータを動的に調整することにより，ユーザの曖昧な検索目標や時間的に変化する検索目標に追従することが可能となる．AIQEC を WWW 情報検索に適用し，基礎的な実験に基づいてその有効性を示す．

2 文書クラスタリングに基づく適応的かつ漸次的なクエリの拡張 (AIQEC)

CBB の過程において文書クラスタに対する適合性評価からユーザの興味を学習するため，従来用いられてきた Rocchio の式を次式のように修正する．

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

ただし，右辺における演算の結果，重みが負の値をとる語についてはその重みを 0 とする．RC と NC はそれぞれ適合クラスタ G_r，不適合クラスタ G_n の集合である．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．α, β はそれぞれ正，負のフィードバックの度合を示すパラメータである．

CBB に適用するため，ARF を文書クラスタリングに基づいて拡張する．まず，適合評価の場合について次の仮定を設ける．例えば，ユーザの検索目標が推移する場合では，クエリと適合文書クラスタ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，適合評価した文書クラスタから得られる情報によるフィードバックを特に強調する．また，例として，ユーザの検索目標が定常的である場合では，クエリと適合文書クラスタが近接することが多いと考える．このとき，α の値は従来用いられてきた 2 に近い値をとり，適合文書クラスタから得られる情報は付加的に利用するにとどめる．なお，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クラスタが近接する場合式 (1) の β は大きくとり，クエリと不適合文書クラスタが近接しない場合は従来用いられてきた 0.5 に近い値をとる．以上のような考えに基づいて，以下に三つの関数を提案する．これらは式 (1) のフィードバックパラメータ α, β の値を与えるものである．

関数 A　文書クラスタの重心ベクトルを用いることによって，フィードバックパラメータを得る．これを関数 A として次式に示す．ただし，$\hat{\mathbf{d}}_i, \hat{\mathbf{d}}_j$ 及び $\hat{\mathbf{q}}_k$ は正規化されている．また，$(p_{k,r}, p_{k,n}) = (a, b)$ において関数は連続である．これらは，後述の関数 B 及び関数 C についても同様である．

$$\alpha = \begin{cases} 1/(c_1^\alpha + c_2^\alpha \cdot p_{k,r}) & (p_{k,r} \le a) \\ 2 & (p_{k,r} > a) \end{cases}, \quad (2)$$

$$\beta = \begin{cases} 0.5 & (p_{k,n} \le b) \\ c_1^\beta + c_2^\beta \cdot p_{k,n} & (p_{k,n} > b) \end{cases}, \quad (3)$$

$$p_{k,r} = \max_{G_r \in RC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{c}}(G_r) \rangle, \quad (4)$$

$$p_{k,n} = \max_{G_n \in NC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{c}}(G_n) \rangle, \quad (5)$$

$$\hat{\mathbf{c}}(G_r) = \mathbf{c}(G_r) / \| \mathbf{c}(G_r) \|, \quad (6)$$

$$\hat{\mathbf{c}}(G_n) = \mathbf{c}(G_n) / \| \mathbf{c}(G_n) \|, \quad (7)$$

$$\mathbf{c}(G_r) = (1/|G_r|) \cdot \sum_{\hat{\mathbf{d}}_i \in G_r} \hat{\mathbf{d}}_i, \quad (8)$$

$$\mathbf{c}(G_n) = (1/|G_n|) \cdot \sum_{\hat{\mathbf{d}}_j \in G_n} \hat{\mathbf{d}}_j. \quad (9)$$

関数 B　適合（不適合）クラスタ内のそれぞれにおいて，前回の検索で用いられたクエリに類似するいくつかの文書の重心ベクトルに基づく値を，式 (2), (3) に代入する．これを関数 B として次式に示す．ただし，$\Lambda(G_r, \hat{\mathbf{q}}_k, m)$ 及び $\Lambda(G_n, \hat{\mathbf{q}}_k, m)$ は，それぞれ $\langle \hat{\mathbf{q}}_k, \hat{\mathbf{d}}_i \rangle$ $(\hat{\mathbf{d}}_i \in G_r)$ 及び $\langle \hat{\mathbf{q}}_k, \hat{\mathbf{d}}_j \rangle$ $(\hat{\mathbf{d}}_j \in G_n)$ の値が大きい順に m 個の文書ベクトルからなる集合を示す．定数 m の値が十分大きいとき，関数 B は関数 A と等価である．

$$p_{k,r} = \max_{G_r \in RC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(G_r) \rangle, \quad (10)$$

$$p_{k,n} = \max_{G_n \in NC} \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(G_n) \rangle, \quad (11)$$

$$\hat{\mathbf{s}}(G_r) = \mathbf{s}(G_r) / \| \mathbf{s}(G_r) \|, \quad (12)$$

$$\hat{\mathbf{s}}(G_n) = \mathbf{s}(G_n) / \| \mathbf{s}(G_n) \|, \quad (13)$$

$$\mathbf{s}(G_r) = (1/m) \cdot \sum_{\hat{\mathbf{d}}_i \in \Lambda(G_r, \hat{\mathbf{q}}_k, m)} \hat{\mathbf{d}}_i, \quad (14)$$

$$\mathbf{s}(G_n) = (1/m) \cdot \sum_{\hat{\mathbf{d}}_j \in \Lambda(G_n, \hat{\mathbf{q}}_k, m)} \hat{\mathbf{d}}_j \quad (15)$$

Adaptive and Incremental Query Expansion Enhanced for Cluster-based Browsing

Koji Eguchi, Hidetaka Ito, Akira Kumamoto and Yakichi Kanata, Faculty of Engineering, Kansai University

表 1: ユーザの検索目標が推移する場合と定常的である場合の平均適合率.

	shifted			constant		
function	(1) average precision	(2) % increase	(3) $\bar{\alpha}$	(4) average precision	(5) % increase	(6) $\bar{\alpha}$
fixed	0.259	0.0	2.0	0.600	0.0	2.0
A	0.412	59.1	11.5	0.650	8.3	2.2
B (m=3)	0.415	60.2	14.1	0.654	9.0	2.5
C (m=3)	0.414	59.8	13.9	0.653	8.8	2.5
C (m=1)	0.395	52.5	10.6	0.653	8.8	2.3

関数C 適合（不適合）クラスタを併合することによって生成される文書集合において，前回の検索で用いられたクエリに類似するいくつかの文書の重心ベクトルに基づく値を，式(2),(3)に代入する．これを関数Cとして次式に示す．ただし，$\cup_{G_r \in RC}$ 及び $\cup_{G_n \in NC}$ は，それぞれ RC における G_r の和集合，NC における G_n の和集合を示す．$|RC| \leq 1$，$|NC| \leq 1$ のとき，関数Cは関数Bと等価である．

$$p_{k,r} = \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(\cup_{G_r \in RC}) \rangle, \quad (16)$$

$$p_{k,n} = \langle \hat{\mathbf{q}}_k, \hat{\mathbf{s}}(\cup_{G_n \in NC}) \rangle, \quad (17)$$

$$\hat{\mathbf{s}}(\cup_{G_r \in RC}) = \mathbf{s}(\cup_{G_r \in RC}) / \| \mathbf{s}(\cup_{G_r \in RC}) \|, \quad (18)$$

$$\hat{\mathbf{s}}(\cup_{G_n \in NC}) = \mathbf{s}(\cup_{G_n \in NC}) / \| \mathbf{s}(\cup_{G_n \in NC}) \|, \quad (19)$$

$$\mathbf{s}(\cup_{G_r \in RC}) = (1/m) \cdot \sum_{\hat{\mathbf{d}}_i \in A(\cup_{G_r \in RC}, \hat{\mathbf{q}}_k, m)} \hat{\mathbf{d}}_i, \quad (20)$$

$$\mathbf{s}(\cup_{G_n \in NC}) = (1/m) \cdot \sum_{\hat{\mathbf{d}}_j \in A(\cup_{G_r \in RC}, \hat{\mathbf{q}}_k, m)} \hat{\mathbf{d}}_j. \quad (21)$$

さて，AIQECにおいては，CBBの過程でクラスタリングされた検索結果に対して，ユーザが適合（不適合）情報を含むと判断したクラスタを選択するたびごとに，検索に用いられたクエリが上記三種のいずれかの関数に基づいて漸次的に拡張・洗練化され，ユーザに提示される．なお，ユーザはクエリから不必要な語を除去することが許可されている．更新されたクエリは随時，再検索に用いることができる．

3 実験及び検討

本実験では，WWWロボットにより収集した一万件の文書集合に対して実験を行う．なお，**2**で述べたAIQECを実現するため，式(2),(3)において $a = 0.679$，$b = 0.339$，$\alpha(p_{k,r} = 0) = 100$ 及び $\beta(p_{k,n} = 1) = 1$ とした．結果として，式(2),(3)における係数は，$c_1^\alpha = 0.010$，$c_2^\alpha = 0.722$，$c_1^\beta = 0.244$，$c_2^\beta = 0.756$ となった[1]．

ユーザの検索目標が推移する場合ついて，実験に基づいて式(1)の α，β を従来のように固定する手法とAIQECにより調整する手法の検索精度を比較する．

実験の手順として，まず八種のクエリ[1]を用いた検索を行う．その後，同じ検索目標のもと適合フィードバックにより再検索を一度行う．次に，ユーザの検索目標が推移したと仮定[2]して，その新たな検索目標のもとに適合クラスタを評価し，それに基づいて漸次的に修正されたクエリにより再検索を行う[3]．このとき，固定の α を用いた場合と，関数A，関数B（$m = 3$），関数C（$m = 3$），関数C（$m = 1$）の四つの関数を用いて動的に α を調整した場合について比較する．それぞれの11点平均適合率（以下，平均適合率と呼ぶ）[4]，パラメータ固定の場合のそれを基準とした増加率及び α の平均値 $\bar{\alpha}$ をそれぞれ表1(1),(2),(3)に示す．表1(2)から，この場合においては，動的に調整されたパラメータによる手法はを固定パラメータによる手法より検索精度を約60%改善できることが確認される．ところで，表1(2)から，関数A・関数B（$m = 3$）・関数C（$m = 3$）が同様な検索精度を示しているのに対して，関数C（$m = 1$）は検索精度がやや低いことがわかる．

次に，定常的な検索目標の場合について，フィードバックパラメータを従来のように固定する手法とAIQECにより調整する手法の検索精度を比較する．前述の実験と同様の手順であるが，終始定常的な検索目標に基づいて再検索を行う．各々の関数に対する平均適合率とその増加率及び $\bar{\alpha}$ をそれぞれ表1(4),(5),(6)に示す．表1(5)から，この場合においては，動的に調整されたパラメータによる手法と固定のパラメータによる手法とで検索精度にそれほど差がないことがわかる．

4 まとめ

本稿では，クラスタに基づくブラウジングにおいて適応的かつ漸次的なクエリの拡張を実現する新たな手法，AIQECを提案した．AIQECをWWW情報検索に適用することによる基礎的な実験の結果，以下のことが明らかになった．(1) 推移する検索目標に対して，AIQECは固定パラメータによる手法より有効にその推移に追従できる．(2) 定常的な検索目標に対して，AIQECは固定パラメータによる手法と同等の検索精度を得ることができる．(3) 良好な検索精度を与える関数は，前回のクエリに類似するいくつかの適合クラスタ内文書の重心に基づくものである．また，AIQECを用いることによって，ユーザの適合性評価に要する負荷が軽減される．今後の課題として，大規模な文書集合に対する詳細な実験による各種関数及びパラメータの最適化，クラスタリングの応答性の改善等が考えられる．

参考文献

[1]江口他: 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, 電学論 (C), 117-C, 11, pp. 1643–1649 (1997).

[2]Hearst, M. A., *et al.*: Reexamining the cluster hypothesis: Scatter/Gather on retrieval results, *Proc. SIGIR'96*, pp. 76–84 (1996).

[3]江口他: 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, 情処研報, FI47-11(NL121-19), pp. 135–142 (1997).

[1](a)「マルチエージェント」, (b)「Neural Network」, (c)「ヒューマンインタフェース」, (d)「マルチメディア」, (e)「画像認識」, (f)「自然言語処理」, (g)「データベース」, (h)「並列計算」を初期クエリとして用いた．

[2]それぞれ，(a) オブジェクト指向，(b) ロボティクス，(c) 対話処理，(d) 画像認識，(e) 自然言語処理，(f) 画像認識，(g) マルチメディア，(h) ヒューマンインタフェースを新たな検索目標とした．

[3]簡単のため不適合評価とユーザによる不必要なクエリ語の除去を省略する．また，検索精度の評価を容易にするため，再検索の結果はクラスタリングせずにフラットに表示する．

[4]適合文書の総数は，固定パラメータ，本実験で用いた四つの関数によるAIQEC及びARFの合計6手法による検索の結果の上位60件において，筆者が一定の基準に基づいて個々について適合と判断した文書の和集合をとり，その要素数をもって近似した．