

クラスタにおける選択的情報に基づくクエリ拡張

5K-2

江口 浩二 神門 典子

学術情報センター 研究開発部

1 はじめに

著者は、ユーザの検索目標が漠然としていたり、時間的に推移するような場合に、ユーザと検索システムとのインタラクションを契機とし、直ちにユーザの興味を取り込むことによって、所望の情報を獲得することを支援する枠組を提案してきた[1]。自動的にクラスタリングされた検索結果に対して、ユーザが適合と判断したクラスタの情報に基づき適応的にクエリを拡張する。本稿では、より正確にユーザの興味をクエリへ取り込むことを目的として、適合クラスタにおいて自動的に選択された文書の情報に基づいた、適合フィードバックについて検討する。

2 クラスタに基づく適応的クエリ拡張

著者は、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判断が要求される問題等に対処するため、適応パラメータを有する適合フィードバック (Adaptive Relevance Feedback: ARF) [2] とクラスタに基づくブラウジング (Cluster-Based Browsing: CBB) [3] 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拡張 (Adaptive and Incremental Query Expansion based on document Clustering: AIQEC) を提案してきた[1]。CBB は、(1) システムが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を実行し、(2) ユーザが適合と判断したクラスタ (以下、適合クラスタ) 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。

CBB の過程において文書クラスタに対する適合性判断からユーザの興味を学習するため、従来用いられていた Rocchio の式 [4] を次式のように修正する。

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

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。本研究では、適合とも不適合とも判断されない文書に関してはフィードバックの対象としない。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。

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と類似するいくつかの文

書の重心ベクトルを用いて、式 (2), (3) に代入する。

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

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

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

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

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

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

$$\mathbf{s}(G_r) = (1/m) \cdot \sum_{\hat{\mathbf{d}}_i \in \Lambda(G_r, \hat{\mathbf{q}}_k, m)} \hat{\mathbf{d}}_i, \quad (8)$$

$$\mathbf{s}(G_n) = (1/m) \cdot \sum_{\hat{\mathbf{d}}_j \in \Lambda(G_n, \hat{\mathbf{q}}_k, m)} \hat{\mathbf{d}}_j \quad (9)$$

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。

上式を用いることにより、ユーザの検索目標が推移する場合にクエリと適合文書クラスタが近接しない傾向があることを反映して、式 (1) の α は通常より大きな値となり、適合クラスタ情報によるフィードバックが特に強調される。一方、ユーザの検索目標が一定である場合ではクエリと適合クラスタが近接することが多い傾向を反映して、 α の値は従来用いられてきた 2 に近い値となり、適合文書クラスタから得られる情報は付加的に利用されるにとどまる。

3 クラスタにおける選択的文書情報に基づくクエリ拡張

2節で述べた AIQEC においては、式 (1) に示されている通り、単に適合クラスタに含まれる全ての文書をフィードバックの対象としている。本稿では、時々刻々と推移するユーザの興味を正確に取り込むことを目的とし、フィードバックのパラメータを調整するだけでなく、自動的に選択された最もクエリに近いいくつかの文書のみをフィードバックに利用することを新たに検討する。本稿では簡単のため、負のフィードバックについては省略する。

$$\mathbf{q}_{k+1} = \hat{\mathbf{q}}_k + \frac{\alpha}{n|RC|} \sum_{G_r \in RC} \sum_{\hat{\mathbf{d}}_i \in \Lambda(G_r, \hat{\mathbf{q}}_k, n)} \hat{\mathbf{d}}_i \quad (10)$$

$\Lambda(G_r, \hat{\mathbf{q}}_k, n)$ は式 (8), (9) と同様、 $\langle \hat{\mathbf{q}}_k, \hat{\mathbf{d}}_i \rangle$ ($\hat{\mathbf{d}}_i \in G_r$) の値が大きい順に n 個の文書ベクトルからなる集合を示す。上式における α は式 (2) を用いて計算される。

ユーザが適合情報を含むと判断した検索結果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式 (10), (2) によって表される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。

表 1: ユーザの検索目標が一定である場合と推移する場合の非補間平均適合率.

	constant			shifted		
	(1) average precision	(2) % increase	(3) $\bar{\alpha}$	(4) average precision	(5) % increase	(6) $\bar{\alpha}$
(A) $\alpha = 2.0, n = G_r $	0.551	0.0	2.0	0.268	0.0	2.0
(B) $m = 10, n = G_r $	0.553	0.4	3.2	0.366	36.6	22.5
(C) $m = 10, n = 10$	0.592	7.4	3.2	0.357	33.2	22.5
(D) $m = 3, n = G_r $	0.551	0.0	2.5	0.370	38.1	19.2
(E) $m = 3, n = 3$	0.565	2.5	2.5	0.345	28.7	19.2

4 実験及び検討

本稿では, WWW ロボットにより収集した, 1 万件の日本語 HTML 文書集合に対して実験を行った. また, (2), (3) において $a = 0.679$, $b = 0.339$, $\alpha(p_{k,r} = 0) = 100$ 及び $\beta(p_{k,r} = 1) = 1$ 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.

4.1 一定の検索目標を仮定した実験

まず, ユーザの検索目標が一定である場合について, 提案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を用いた初期検索を行う. 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, 更に同じ検索目標に基づいて提案手法により再検索を実行する. このとき, (A) $\alpha = 2.0, n = |G_r|$, (B) $m = 10, n = |G_r|$, (C) $m = 10, n = 10$, (D) $m = 3, n = |G_r|$, (E) $m = 3, n = 3$ の五通りについて比較する. それぞれの非補間平均適合率 (以下, 平均適合率と呼ぶ)², (A) の場合を基準とした増加率, 及び, α のそれぞれについての八種の検索課題の平均値を, 表 1(1), (2), (3) に示す.

表 1(2) からわかる通り, ユーザの検索目標が一定であることを仮定した場合は, クラスタにおける全ての文書をフィードバックに用いるよりも, クエリに類似するいくつかの文書を利用した方が検索精度を改善できる. その理由は, クラスタに含まれる非適合文書をフィードバックの対象とするのを避けることができるためと考えられる. ただし, フィードバックに用いる文書数は適切な値に設定する必要がある.

4.2 検索目標の推移を仮定した実験

次に, 4.1 項と同様の手順で初期検索, 適合フィードバックによる再検索を一回行なった後, ユーザの検索目標が推移したと仮定して, その新たな検索目標の

もとに判断した適合クラスタに基づいて提案手法によるクエリ拡張及び再検索を実行する. 4.1 項で述べた (A), (B), (C), (D), (E) の五通りについて, 平均適合率, (A) の場合を基準とした増加率, 及び, α のそれぞれの平均値を表 1(4), (5), (6) に示す.

表 1(5) は, ユーザの検索目標が推移することを仮定した場合は, クエリに類似するいくつかの文書をフィードバックするよりも, クラスタにおける多数の文書を用いる方がユーザの興味の推移に適応できることを示唆する. 各文書が扱う話題に従って十分正確にクラスタリングされることは現実的には容易でないが, そのような場合, クラスタの特徴を十分に反映した文書群は必ずしもクエリに類似する文書群と一致せず, 前者を用いてフィードバックを行なう方が良い結果をもたらすことが考えられる.

5 まとめ

本稿では, 検索結果のクラスタに対して, ユーザが適合と判断したクラスタの情報に基づき適応的にクエリを拡張する枠組において, 適合クラスタにおいて自動選択された文書情報に基づくクエリ拡張を提案した. 基礎的な実験により, ユーザの検索目標が一定である場合とそうでない場合によって, フィードバックに適切な文書数が異なることを示唆する結果を得た. クエリと適合クラスタとの距離を手がかりに適切な文書数の閾値を自動調整するなどの拡張が必要であろう. 今後, 更に詳細な実験を実施するとともに, ユーザの興味の変遷をも勘案したインタラクティブ情報検索の評価法について検討する.

参考文献

- [1] Eguchi, K., et al.: Adaptive Query Expansion Based on Clustering Search Results, 情処学論, 40, 5, pp. 2439-2449 (1999).
- [2] 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).
- [3] Hearst, M. A., et al.: Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, Proc. ACM SIGIR '96, pp. 76-84 (1996).
- [4] 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).

¹(a1)「マルチエージェント」, (b1)「Neural Network」, (c1)「ヒューマンインタフェース」, (d1)「マルチメディア」, (e1)「画像認識」, (f1)「自然言語処理」, (g1)「データベース」, (h1)「並列計算」を初期クエリとして用いた. 4.2 項における推移の例としては, それぞれに対して, (a2) オブジェクト指向, (b2) ロボティクス, (c2) 対話処理, (d2) 画像認識, (e2) 自然言語処理, (f2) 画像認識, (g2) マルチメディア, (h2) ヒューマンインタフェースを新たな検索目標とした.

²適合文書の総数は, 本実験で用いた五種の組合せのパラメータを用いた検索, 及び, クラスタリングを伴わない手法である ARF[2] による検索の各々の結果の上位 100 件において, 筆頭著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. 簡単のため不適合判断とユーザによる不必要なクエリ語の除去を省略した. 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を行なわなかった.