D-044

クラスタに基づいた適合性フィードバック手法
Relevance Feedback Methods Based on Clusters

湯木野 高幸 † 松下 大輔 † 平澤 茂一 †
Takayuki Yukino Daisuke Matsushita Shigeichi Hirasawa

1. はじめに

今日，インターネットの普及により個人で扱うことができるテキストデータが氾濫しており，その大規模なデータ群から必要な文書を効率的に探し出す情報検索技術は強力なツールとなる．

情報検索技術において，代表的なモデルにベクトル空間モデルがある。ベクトル空間モデルにおいて，ユーザとの対話によって検索質問を拡張し，再検索を行う適合性フィードバックに基づいた手法が提案されており，検索システムの精度を向上することが知られている [2]．従来の手法は検索結果の各々の文書についてユーザが適合 · 不適合の判定をする必要がある．しかし，ユーザへの負担を考慮し，検索システムは極力少ない文書を基にフィードバックをすることが望ましい．

このような背景から，本研究は検索システムが出力した文書集合に対してクラスタリングを行い，フィードバックすることにより，より少ないフィードバック文書においても検索精度を向上させる方法を提案する．また，テスト文書セット [5][6] に適応し，有効性を示す．

2. 従来手法

ベクトル空間モデルにおける検索では，形態素解析処理により文書集合から索引語を抽出し，この索引語の重みを要素とするベクトルで文書を表現する [2]．検索質問もまた索引語の重みを要素とするクエリベクトルで表現し，クエリベクトルと文書ベクトルの類似度を計算する．類似度の降順に検索結果 (ランキング) を出力する．

[定義 1: 索引語の重み $w_{d_j}^{t_k}$]

索引語の重み $w_{d_j}^{t_k}$ は以下の TF · IDF 値で与えられる．

$$w_{d_j}^{t_k} = (f(t_k, d_j)/F(d_j)) \times (1 + \log(M/df(t_k))). \quad (1)$$

d_j : 検索対象文書 $(j = 1, 2, \ldots, M)$
t_k : 検索対象文書集合に出現する単語 $(k = 1, \ldots, N)$
$f(t_k, d_j)$: 文書 d_j における単語 t_k の出現回数
$F(d_j)$: 文書 d_j の全単語数
$df(t_k)$: 単語 t_k が出現する文書数

[定義 2: 文書ベクトル d_j]

文書 d_j の文書ベクトル $\boldsymbol{d_j}$ を次式で与える．

$$\boldsymbol{d_j} = (w_{d_j}^{t_1}, w_{d_j}^{t_2}, \ldots, w_{d_j}^{t_N}). \quad (2)$$

[定義 3: クエリベクトル Q]

検索質問は次式のクエリベクトルで表される．

$$\boldsymbol{Q} = (q^{t_1}, q^{t_2}, \ldots, q^{t_N}). \quad (3)$$

$$q^{t_k} = \begin{cases} 0 & \text{単語 } t_k \text{ が検索キーワードでない} \\ 1 & \text{単語 } t_k \text{ が検索キーワードである} \end{cases}$$

[定義 4: ベクトル間の類似度 $\sigma(\boldsymbol{x},\boldsymbol{y})$]

ベクトル$\boldsymbol{x}$，$\boldsymbol{y}$の類似度は以下 $\sigma(\boldsymbol{x},\boldsymbol{y})$ で与えられる．

$$\sigma(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x},\boldsymbol{y})/|\boldsymbol{x}||\boldsymbol{y}|. \quad (4)$$

$(\boldsymbol{x},\boldsymbol{y})$ はベクトル $\boldsymbol{x}$ と $\boldsymbol{y}$ の内積，$|\boldsymbol{x}|$ はベクトル $\boldsymbol{x}$ のノルムを示す．

†早稲田大学大学院理工学研究科経営システム工学専攻

2.1 適合性フィードバック

適合フィードバックは，初期検索結果で得た文書集合の一部に対し，ユーザが適合 · 不適合の判定行いその情報をもとに検索質問を更新し，システムに返すことで検索システムの精度を対話的に改善する手法である．

2.2 Rocchio アルゴリズム [1]

[定義 5:Rocchio の検索質問更新式]

$$\begin{aligned} \boldsymbol{Q}_{new} &= (q_{new}^{t_1}, q_{new}^{t_2}, \ldots, q_{new}^{t_N}) \\ &= \alpha\,\boldsymbol{Q} + \frac{\beta}{|D^+|}\sum_{d_j \in D^+} \boldsymbol{d_j} - \frac{\gamma}{|D^-|}\sum_{d_j \in D^-} \boldsymbol{d_j}. \quad (5) \end{aligned}$$

$\alpha, \beta, \gamma (> 0)$：重み係数
$D^+(D^-)$:ユーザが (不) 適合と判断した文書集合
$|D^+|(D^-)$:ユーザが (不) 適合と判断した文書数

3. 提案手法

3.1 従来手法の問題点

従来手法では，フィードバックの際にユーザは検索システムの出力から得られた文書集合のうち，上位数文書を適合文書と不適合文書に判別し，検索システムに返す．はじめに述べたように，ユーザの負担を考慮し, 少ないフィードバック文書をユーザに判定させるべきである．

3.2 提案手法の目的と概要

提案手法では出力された文書集合をクラスタリングし，作成したクラスタを内容の類似した文書の集合と捉える。各クラスタから代表文書を抽出し，ユーザに提示し，判定させることでユーザに負担を軽減する．

3.3 概念ベクトル [3]

文書集合をいくつかのクラスタに分けることにより，内容の類似したクラスタ集合を得ることができる．クラスタの重心ベクトルを概念ベクトルと呼び，これはクラスタの内容を表す代表ベクトルである．正規化された n 個の文書ベクトルを$\boldsymbol{x}$,$\boldsymbol{x_2}$,$\cdots$,$\boldsymbol{x_n}$とし，$\pi_1\pi_2\cdots\pi_k$ を k 個のクラスタ集合とする．π_j の重心ベクトルを$\boldsymbol{m_j}$とすると，

$$\boldsymbol{m_j} = \frac{1}{n_j}\sum_{\boldsymbol{x}\in\pi_j} \boldsymbol{x} \quad (6)$$

概念ベクトル$\boldsymbol{c_j}$は，重心ベクトル$\boldsymbol{m_j}$を正規化することにより求まる．

[定義 6:概念ベクトル]

$$\boldsymbol{c_j} = \frac{\boldsymbol{m_j}}{||\boldsymbol{m_j}||} \quad (7)$$

3.4 球面 k 平均法

球面 k 平均法により，次の目的関数を最大にするクラスタ集合を作成する．

$$Q(\{\pi_j\}_{j=1}^k) = \sum_{j=1}^{k}\sum_{\boldsymbol{x}\in\pi_j} \boldsymbol{x}^T\boldsymbol{c_j} \quad (8)$$

[アルゴリズム]

S1) フィードバック対象文書集合から k 個のクラスタ $\{\pi_j^{(0)}\}_{j=1}^k$ を得る．概念ベクトルを $\{\boldsymbol{c}_j^{(0)}\}_{j=1}^k$ とする．

S2) 各文書ベクトル $\boldsymbol{x}_i(1 \leq i \leq n)$ に対し，$\boldsymbol{x}_i$ との余弦がもっとも大きい (近い) 概念ベクトルを見つけ，新たな部分集合 $\{\pi_j^{(t+1)}\}_{j=1}^{k}$ を得る．

$$\pi_j^{t+1} = \{\boldsymbol{x} \in \{\boldsymbol{x}_j\}_{i=1}^{n} : \boldsymbol{x}^T \mathrm{c}_j^{(t)} \geq \boldsymbol{x}^T \mathrm{c}_l^{(t)}, 1 \leq l \leq n\} \quad (9)$$

$(1 \leq j \leq k)$

S3) 新たに得られたクラスタから重心を計算し，正規化することで概念ベクトルを得る．

$$\boldsymbol{c}_j^{(t+1)} = \frac{\boldsymbol{m}_j^{(t+1)}}{||\boldsymbol{m}_j^{(t+1)}||}(1 \leq j \leq k) \quad (10)$$

S4) 停止基準を満たすと，$\pi_j^{\star} = \pi_j^{t+1}, \mathrm{c}_j^{\star} = \mathrm{c}_j^{t+1}(1 \leq j \leq k)$ となり，アルゴリズムは終了する．停止基準を満たさなければ，$t \rightarrow t+1$ として S2 に戻る．

停止条件，

$$|Q(\{\pi_j^{(t)}\}_{j=1}^{k}) - Q(\{\pi_j^{(t+1)}\}_{j=1}^{k})| \leq \varepsilon \quad (11)$$

3.5 代表文書選定方法

上記のアルゴリズムによって得られたクラスタの中からユーザに提示する代表文書を決定する。提案手法では、概念ベクトルに最も近い文書ベクトルがそのクラスタの内容を最もよく表すベクトルであると考え、そのクラスタの代表文書に選定する。

3.6 適合クラスタからの検索方法

林下ら [4] は，正解文書集合が散在していると仮定し，複数のクエリベクトルを用いてフィードバックすることにより精度の高い検索を実現した．本研究では，クラスタリングによって作成された部分文書集合はそれぞれ異なる概念を表し，それが概念ベクトルに表れると考える．そこで，適合クラスタの概念ベクトル近辺には適合文書が密集すると考え，これに近い文書が適合文書になるという仮定の下に再検索する．検索対象文書の類似度は以下で定義する．

[定義 7:検索対象文書 d_j と適合クラスタ Q との類似度 $Sim'(\boldsymbol{d}_j, Q)$]

$$Sim(\boldsymbol{d}_j, Q) = \max_{\boldsymbol{c}_i \in Q} Sim(\boldsymbol{d}_j, \boldsymbol{c}_i) \quad (12)$$

3.7 提案手法のアルゴリズム

S1) 初期検索により得た，ランキング上位文書 L 個をフィードバック文書集合とする．

S2) ランキング上位文書集合から，球形 k 平均法により k 個の部分集合を作成する．

S3) 作成された各々のクラスタの概念ベクトルから代表文書を得る．

S4) ユーザに代表文書を提示し，ユーザの判別により適合クラスタ・不適合クラスタを得る．

S5) 得られた適合クラスタの概念ベクトル $\boldsymbol{c}_i$ により再検索を行う．

4. シミュレーションと考察

4.1 評価方法

提案手法の有効性を検証するため，シミュレーションによる実験を行った．評価方法としては，再現率，適合率に対し，検索課題ごとの再現率 $0.0, 0.1, \ldots, 1.0$ における適合率（11 点適合率）とその平均値（平均適合率）を用いた．

$$\text{再現率} = \frac{\text{検索された正解文書数}}{\text{正解文書数}}, \quad (13)$$

$$\text{適合率} = \frac{\text{検索された正解文書数}}{\text{検索された総文書数}}. \quad (14)$$

4.2 シミュレーション条件

評価データとして毎日新聞 1994 [5] をもとにした BMIR-J2 テストコレクション（5,080 文書）[6] を用いた．また評価対象 9 課題に対するユーザの適合・不適合の判定には BMIR-J2 が提供する正解文書，不正解文書を使用した．式 (11) において $\varepsilon = 10^{-8}$ とした．また，従来手法に用いられる式 (5) のパラメータの値に関しては，経験的に精度が高いとされる， α =1, β =1, γ =0.5 とした．

4.3 結果

図 1 には，フィードバック文書数 $L = 30$ においてクラスタ (部分集合) 数を変化させた場合の 11 点適合率について算出し，平均をとった再現率・適合率曲線を示す．また，表 1 には全課題に対する初期検索，従来手法による検索，提案手法における検索の 11 点適合率の平均を示す．

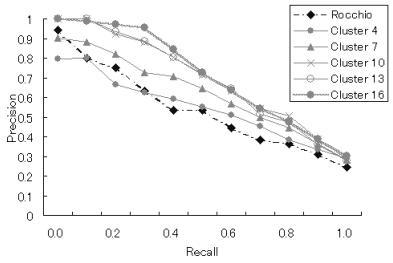

図 1: 再現率・適合率曲線

	初期	従来
平均適合率	0.488	0.541

クラスタ数 k	4	7	10	12	15
平均適合率	0.545	0.622	0.698	0.704	0.712

4.4 考察

1. 図 1 によると，提案手法はクラスタ数が減少するほど検索精度が低下する．特に，クラスタ数が 5 以上の場合に従来手法の検索精度を上回ることになる．提案手法は 5 文書をユーザに提示することで，30 文書をユーザに提示する従来手法と同精度以上の検索を行うことができる．
2. 他の実験でクラスタ数を固定し，フィードバック文書を増加させても検索精度を向上することができなかった．これらの事実から，本手法はフィードバック文書数よりクラスタ数に影響を受けていることがわかる．
3. 本手法はクラスタ数が少ない場合 (ユーザに提示する文書数) においても高い検索精度で検索が可能であり，有効な手法である．

5. まとめと今後の課題

クラスタによる適合性フィードバックは検索精度を上げる，またはユーザの負担を軽減することが可能な手法であることを示した．今後は，自動的にクラスタ数 k を決定できる手法を検討していく予定である．

参考文献

[1] 北研二, 津田和彦, 獅子堀正幹, 情報検索アルゴリズム, 共立出版,1999 年.

[2] 徳永健伸, 情報検索と言語処理, 財団法人東京大学出版会,1999 年.

[3] Dhillon, I., and D. Modha.,*Concept Decompositions for Large Sparse Text Data using Clustering*,Technical Report RJ 10147 (95022). IBM Almaden Research Center,1999.

[4] 林下雄也, 八木秀樹, 平澤茂一"複数のクエリベクトルを用いた適合税フィードバック手法", 第 27 回情報理論とその応用シンポジウム,Vol.1,pp.49-pp.52,2004.

[5] 毎日新聞社,CD 毎日新聞'94, 日外アソシエーツ, 1995 年.

[6] (社) 情報処理学会データベースシステム研究会, BMIR-J2, 新情報処理開発機構, 1998 年.