The 29th Symposium on Information Theory
and its Applications (SITA2006)
Hakodate, Hokkaido, Japan, Nov. 28 – Dec. 1, 2006

階層的クラスタを用いた適合性フィードバック手法による文書検索
Document Retrieval based on Relevance Feedback Using Hierarchical Cluster

穴瀬 裕之* 石田 崇* 平澤 茂一*
Hiroyuki ANASE Takashi ISHIDA Shigeichi HIRASAWA

Abstract— Generally, user's interest is wide-ranged and unclear. In the information retrieval systems, it is difficult for a user to input his or her query with suitable keywords. In this paper, we propose a new document retrieval method based on relevance feedback using hierarchical cluster. The proposed method creates hierarchical cluster from a document set. By using it, query expansion with user feedback information works more effectively. Finally, we show that this method improves the retrieval accuracy compared with the conventional relevance feedback by numerical experiment.

Keywords— relevance feedback, hierarchical cluster, Rocchio algorithm

1 はじめに

情報検索において，一般にユーザの関心は広範かつあいまいであり，ユーザが適切な検索質問を作成することは困難であることが知られている．この問題を解消するための検索質問拡張の手法として，ユーザに初期検索結果の適合・不適合の判定をしてもらうことで情報検索の精度を向上させる適合性フィードバックの研究が盛んに行われている [1].

しかしベクトル空間モデルに基づく情報検索では所望の文書は空間上に散在しており，ユーザから得られたフィードバック情報をそのまま用いるだけでは充分に機能しないという問題点が挙げられる．

本研究ではこの問題を解消すべく，対象文書集合は階層構造を成しているという仮定のもとに，階層的クラスタリングを用いた検索質問拡張手法を提案し，検索精度の向上を図る．この手法を情報検索のベンチマークデータである BMIR-J2[2][3] に適用し実験的に有効性を示す．

2 準備

本研究では，最も一般的な検索モデルであるベクトル空間モデル (VSM)[1] を用いる．

2.1 ベクトル空間モデル

ベクトル空間モデルでは，文書とユーザの検索質問をベクトルとして表現する．これによって，各文書がどれくらい検索質問に適しているかをベクトル間の類似度に帰着させることができる．このベクトル空間は単語ごとに独立した次元を持ち，文書は単語の重みを要素とするベクトルとして表される．

ベクトルにおける単語の重みを用いて計算される TF·IDF 法 [4] では，文書データベース中の多くの文書に出現する単語は重要ではなく，特定の文書において多く出現する単語は重要とすることで単語の重みを決定するものである．また，ベクトル同士の類似度は両ベクトルの余弦によって計算する．文書を $d_j(j=1,2,\cdots,M)$ とし，単語を $t_i(i=1,2,\cdots,n)$ と表す．

定義 1: (文書 d_j 中の単語 t_i の TF·IDF 値：w_{ij})

$$w_{ij} = \frac{tf(t_i,d_j)}{\sum_{t\in d_j} tf(t,d_j)} \cdot \left(\log_{10}\frac{M}{df(t_i)}+1\right) \qquad (1)$$

$tf(t_i,d_j)$：文書 d_j 内の単語 t_i の出現頻度
$df(t_i)$：単語 t_i 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 $\boldsymbol{d_j}$ は TF·IDF 値により以下のように与えられる．

$$\boldsymbol{d_j} = (w_{1j}, w_{2j}, \ldots, w_{nj}) \qquad (2)$$

n:システム中の全単語数

定義 3: (検索質問ベクトル q)
検索質問 q は次式で定義する検索質問ベクトル $\boldsymbol{q}$ として表される．

$$\boldsymbol{q} = (q^1, q^2, \ldots, q^n) \qquad (3)$$

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

定義 4: (d_j と q の類似度)
ベクトル d_j, q の類似度は以下の式で与えられる．

$$sim(d_j,q) = \frac{(\boldsymbol{d_j},\boldsymbol{q})}{|\boldsymbol{d_j}||\boldsymbol{q}|} \qquad (4)$$

$(\boldsymbol{d_j},\boldsymbol{q})$ はベクトル d_j と q の内積，$|\boldsymbol{d_j}|$, $|\boldsymbol{q}|$ はそれぞれベクトル d_j, q のノルムを示す．

* 〒 169–8555 東京都新宿区大久保 3–4–1 早稲田大学理工学部経営システム工学科 School of Science and Engineering Waseda University 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan. E-mail: anase@hirasa.mgmt.waseda.ac.jp

2.2 適合性フィードバック

適合性フィードバックは，初期検索結果で得た文書集合のうち一部の上位文書に対しユーザが適合・不適合の判定を行い，検索システムがその情報を用いて検索質問を更新することで検索精度を対話的に改善する手法である．

代表的な適合性フィードバック手法のひとつに，Rocchio アルゴリズム [5] がある．Rocchio アルゴリズムでは，元の検索質問ベクトル q に適合文書集合の重心ベクトルと，不適合文書集合の重心ベクトルとの差分ベクトルを加算することにより検索質問ベクトルを更新して q_{new} とし，再度検索を行う手法である．

定義 5: (Rocchio の検索質問更新式)

$$q_{new} = (q_{new}^1, q_{new}^2, \ldots, q_{new}^n) = \alpha\, q + \frac{\beta}{|D^+|} \sum_{d_j \in D^+} d_j - \frac{\gamma}{|D^-|} \sum_{d_j \in D^-} d_j. \quad (5)$$

$\alpha, \beta, \gamma (> 0)$：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

2.3 クラスタリング手法

本節では提案手法に用いるクラスタ分析の代表的な手法である，球面 k-means 法 [6] 及び凝集法 [7] について述べる．

2.3.1 球面 k-means 法

球面 k-means 法 [6] とは，代表的な非階層的クラスタリング手法のひとつである．以下に文書集合に対する球面 k-means 法アルゴリズムを示す．

[球面 k-means 法アルゴリズム]

S0) ベクトル空間上にランダムに k 個のベクトルを作成し，これを k 個の初期クラスタ π_j の重心ベクトル c_j とする．また以下の式を目的関数とする．

$$\arg\max_{\{\pi_j\}_{j=1}^k} Q(\{\pi_j\}_{j=1}^k) = \sum_{j=1}^{k} \sum_{d \in \pi_j} (d \cdot c_j) \quad (6)$$

S1) 検索対象文書集合から k 個のクラスタ $\{\pi_j^{(0)}\}_{j=1}^k$ を得る．またその重心ベクトルを $\{c_j^{(0)}\}_{j=1}^k$ とする．

S2) 各文書ベクトル $d_i (1 \le i \le n)$ に対し，d_i と類似度の最も高い（余弦がもっとも大きい）重心ベクトルを探し，新たな部分集合 $\{\pi_j^{(t+1)}\}_{j=1}^k$ を得る．

$$\pi_j^{t+1} = \{d \in \{d_j\}_{i=1}^n : (d, c_j^{(t)}) \ge (d, c_l^{(t)}), 1 \le l \le n\} \quad (7)$$

$(1 \le j \le k)$

S3) 新たに得られたクラスタの重心を計算し，正規化することで各重心ベクトルを更新する．

$$m_j = \frac{1}{n_j} \sum_{d \in \pi_j} d \quad (8)$$

$$c_j^{(t+1)} = \frac{m_j^{(t+1)}}{||m_j^{(t+1)}||} (1 \le j \le k) \quad (9)$$

S4) 以下の停止基準を満たすと，$\pi_j^* = \pi_j^{t+1}, c_j^* = c_j^{t+1} (1 \le j \le k)$ となり，アルゴリズムは終了．目的関数最大となるクラスタ集合が生成される．停止基準を満たさなければ，$t \to t+1$ として S2) に戻る．

$$|Q(\{\pi_j^{(t)}\}_{j=1}^k) - Q(\{\pi_j^{(t+1)}\}_{j=1}^k)| \le \varepsilon \quad (10)$$
$$(\varepsilon \ge 0)$$

□

2.3.2 凝集法

凝集法は代表的な階層的クラスタリング手法であり，局所的な類似性を表現するのに適している．以下で文書集合に対する凝集法アルゴリズムを示す．

[凝集法アルゴリズム]

S1) N 個の文書について 1 個の文書を 1 個のクラスタとして，それぞれのクラスタ間の類似度を算出し，類似度行列を作成する．また類似度にはクラスタの重心ベクトル同士の余弦を用いる．

S2) 類似度が最大となる 2 つのクラスタを併合し，1 つのクラスタとする．

S3) 併合後のクラスタと他のクラスタとの類似度を計算し，類似度行列を更新する．

S4) クラスタ数が 1 になれば終了．停止基準を満たさなければ S2) へ戻る． □

3 提案手法

3.1 従来手法の問題点

Rocchio アルゴリズム [5] では，ある検索質問に対する適合文書集合は密集しており適合文書ベクトルの重心と検索質問ベクトルが重なる場合に極めて良好な検索精度を達成すると仮定し，これに近づくような検索質問更新を行う手法である．しかし林下らは，適合文書は局所的に密集しているという仮定に基づき Rocchio アルゴリズムよりも大きく上回る検索性能を達成した [9]．林下らの仮定のように，適合文書が散在しているのであれば Rocchio アルゴリズムのように単純に重心ベクトルを加算するだけでは適切な検索質問更新が行われない．本論では林下らの仮定に基づき議論を進める．

3.2 提案手法の概要

提案手法では，「文書が類似していれば，同じ検索質問に対する適合性も同様に類似している」というクラスタ

仮説 [8] に基づき，まずは球面 k-means 法を用いて文書をクラスタリングし，文書集合を大域的に分割する．次に，林下らの仮説に基づき提案手法では局所的な適合文書集合を表現するため，球面 k-means 法により得られた k 個のクラスタ $c_1, c_2, \cdots, c_k$ に対し凝集法によるクラスタリングを行い文書集合をクラスタ階層構造として表現する．階層構造を用いることで非常に局所的な文書の類似性を捉えることができ，適合文書と類似した文書を容易に見つけ出すことができる．図 1 に生成された階層的クラスタを示す．

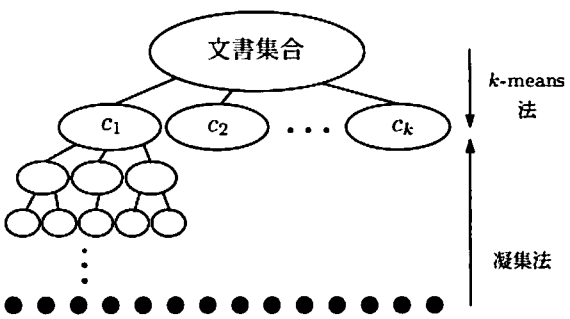

図 1: 階層的クラスタの概要

3.3 検索質問拡張

本研究での狙いは，散在した適合文書集合を的確に表現した階層的クラスタを用いて，フィードバックで得られた適合文書と類似した文書を同様に適合文書と推定し，散在した適合文書集合をより的確に収集することである．そこで提案手法では，適合文書を用いて散在する文書集合をクラスタとして抽出する．

[適合クラスタの抽出アルゴリズム]

S0) 適合文書 $d_i(1 \leq i \leq R)$ に対し初期値として $i = 1$ を与える．(R:フィードバックにおける適合文書数)

S1) d_i が所属するクラスタ $c_m(1 \leq m \leq k)$ を選択する．

S2) 選択したクラスタが条件 1 ～ 3 を満たすなら，そのクラスタを適合クラスタ c_i として保存．そうでなければ現在のクラスタの子のうち d_i が含まれるクラスタへ移動．

条件 1) クラスタ内に不適合文書が含まれない
条件 2) クラスタ内に他の適合文書が含まれない
条件 3) クラスタ内文書数が x 未満である

S3) $i = R$ にならば終了．そうでなければ $i \rightarrow i + 1$ として S1 へ． □

次に，不適合文書についても同様の操作を行う．

[不適合クラスタの抽出アルゴリズム]

S0) 不適合文書 $d_j(1 \leq j \leq I)$ に対し初期値として $j = 1$ を与える．(I:フィードバックにおける不適合文書数)

S1) d_i が所属するクラスタ $c_m(1 \leq m \leq k)$ を選択する．

S2) 選択したクラスタが条件 1 ～ 3 を満たすならそのクラスタを不適合クラスタ c_j として保存．そうでなければ現在のクラスタの子のうち d_j が含まれるクラスタへ移動．

条件 1) クラスタ内に適合文書が含まれない
条件 2) クラスタ内に他の不適合文書が含まれない
条件 3) クラスタ内文書数が x 未満である

S3) $j = I$ ならば終了．そうでなければ $j \rightarrow j + 1$ として S1 へ． □

ここで得られた適合クラスタの重心ベクトルは散在し局所で密集している適合文書の概念を表すと考えられる．また不適合文書についてはクラスタ仮説に基づき，不適合文書と極めて類似する文書もまた不適合文書であると考えて擬似的に不適合文書を推定し，検索質問ベクトル更新に実際の不適合文書よりも多くの文書を用いる．これらの重心を用いて以下の式により検索質問ベクトルを更新する．

$$q_{new} = \alpha\, q + \frac{\beta}{|C^+|} \sum_{c_j \in C^+} c_j - \frac{\gamma}{|C^-|} \sum_{c_j \in C^-} c_j. \quad (11)$$

$\alpha, \beta, \gamma (> 0)$：重み係数
$C^+(C^-)$:(不) 適合クラスタ集合
$|C^+|(|C^-|)$:(不) 適合クラスタ数

また両アルゴリズムにおける条件 3 は，適合・不適合文書との類似度が非常に高い文書以外を除去する目的で用いた．

3.4 提案アルゴリズム

以下に提案アルゴリズムを示す．

S1) ユーザは検索質問を入力する．

S2) 検索システムは文書集合と検索質問の類似度を計算し，類似度順に並べ初期検索結果としてユーザに提示する．

S3) ユーザは提示された上位数文書に対し適合・不適合の判定を行いシステムに返す．

S4) 検索システムはフィードバック情報に対し 3.3 節のアルゴリズムを適用し，適合・不適合クラスタを得る．

S5) 検索質問更新を行い，新たな検索結果をユーザに提示する． □

4 数値実験と考察

4.1 実験条件

評価データとして毎日新聞1994 [2] をもとにしたBMIR-J2テストコレクション（5.080文書，25471単語）[3] を用いた．また評価対象10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．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 [5]．さらに球面 k-means 法のクラスタ数 k =16，3.3節で示したアルゴリズム中の条件3でのクラスタ文書数を $x = 10$ とした．

4.2 評価

従来手法であるRocchioフィードバックと提案手法の検索精度を比較する．評価尺度には，検索結果の適合率と再現率を用いる．さらに，再現率が $0.0, 0.1. 0.2. \cdots, 1.0$ のときの適合率である11点適合率を求め，その適合率の平均である11点平均適合率を算出する．適合率及び再現率を以下に示す．

$$\text{再現率} = \frac{\text{検索された正解文書数}}{\text{正解文書数}} \tag{12}$$

$$\text{適合率} = \frac{\text{検索された正解文書数}}{\text{検索された総文書数}} \tag{13}$$

4.3 結果と考察

フィードバック (FB) 文書数 10,20.30 のときの実験結果を表1に示す．

表 1: フィードバック文書数 10.20,30 のときの実験結果

FB 文書数	初期検索	Rocchio	提案手法
10	0.4454	0.5213	0.5374
20	0.4454	0.5280	0.5361
30	0.4454	0.5284	0.5348

表1より提案手法はどのフィードバック文書数においてもRocchioアルゴリズムよりも良好な検索性能を示していることが分かる．

次に提案と従来の検索精度の差が最も大きかったフィードバック文書数10のときの11点適合率を図2に示す．図2より，提案手法は従来手法に比べわずかながらではあるが安定的に良好な検索性能を示しており，特に再現率が低いところで差が出ている．また実験より，フィードバック文書があるクラスタ $c_m (0 \leq m \leq k)$ に集中する場合に特に良好な結果を示している．これは，3.3節アルゴリズム中の条件1が大きく働くことで，より検索質問拡張に有効な適合・不適合クラスタを抽出することが出来たためと考えられる．

5 まとめと今後の課題

本研究では，前処理として文書集合に階層的クラスタングを行い局所的に密集している適合文書クラスタを抽出することで，より正確に適合文書集合を推定する手法を提案した．またベンチマークを適用したシミュレーションにより提案手法の有効性を示すことができた．

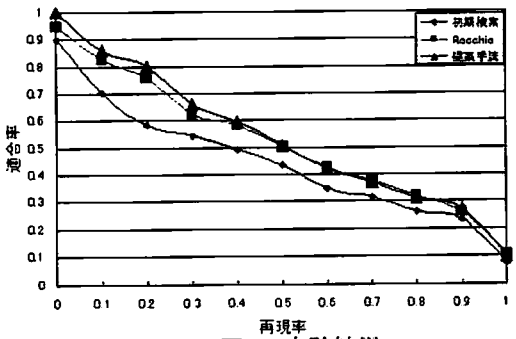

図 2: 実験結果

しかし，提案手法では期待したほどの検索精度は得られなかった．その原因として，複数の文書からなるクラスタの重心を用いるため散在する適合文書クラスタの特徴が薄まってしまうことが考えられる．

今後の課題は，局所的な適合文書の特徴を色濃く表現し収集する方法の検討，たとえば各クラスタから特徴的な単語を抜き出す手法，複数のクエリベクトルを使う手法の導入などが考えられる．

6 謝辞

本研究の成果の一部は，早稲田大学特定課題研究助成2005B-189による．

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] 毎日新聞社，CD毎日新聞'94，日外アソシエーツ，1995年．
[3] (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．
[4] 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．
[5] J. Rocchio. Relevance Feedback in Information Retrieval. *The SMART Retrieval System Experiments in Automatic Document Processing*, Prentice Hall Inc, 1971.
[6] I. Dhillon, D. Modha. "*Concept Decompositions for Large Sparse Text Data using Clustering*"Technical Report RJ 10147 (95022). IBM Almaden Research Center,1999.
[7] 宮本定明，クラスター分析入門，森北出版株式会社，1999.
[8] van. Rijsbergen, C. J. , "Further experiments with hierarchic clustering in document retrieval, " Information Storage and Retrieval, 10, 1-14, 1974.
[9] 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性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.