The 27th Symposium on Information Theory
and Its Applications (SITA2004)
Gero, Gifu, Japan, Dec. 14-17, 2004

PLSIに基づく適合性フィードバック手法
A Relevance Feedback Method based on the Probabilistic Latent Semantic Indexing

足立 鉱史* 石田 崇* 平澤 茂一*
Hiroshi ADACHI Takashi ISHIDA Shigeichi HIRASAWA

Abstract— Recently, the information retrieval technique based on keyword is mainly used as a method for accessing text databases. However, all information that a user needs cannot be necessarily searched by a single query. Thus, the relevance feedback method for improving the performance is usually utilized, by means that a user returns the information whether some information is matched to his reference purpose to a search engine. On the other hand, since a words-documents matrix is sparse and high-dimensional, the PLSI (Probabilistic Latent Semantic Indexing) which compresses dimensions is indispensable. However, as for the present condition, there are few effective and practical relevance feedback methods based on the PLSI. In this paper, a new relevance feedback method based on the PLSI is proposed. And we show the validity of the method by the simulation.

Keywords— information retrieval, relevance feedback, probabilistic latent semantic indexing, vector space model

1 はじめに

計算機システムの高性能化とネットワーク化に伴い，近年，膨大な電子化された情報が計算機上でアクセス可能になっている．情報検索は，これらの膨大な情報の中から必要な情報を見つけ出すために必要不可欠な技術である [1]．情報検索において，一回のみの検索でユーザの必要とする情報をすべて検索できるとは限らない．そこでユーザが，得られた検索結果のうち，どの情報が検索意図に適合し，どの情報が適合しないかを検索システムに教えることにより，システムの検索精度を改善する適合性フィードバック手法がある [1]．

一方，情報検索の分野において，確率論に基づき索引語文書行列を低次元に圧縮するPLSI(Probabilistic Latent Semantic Indexing) と呼ばれる技術が提案され，情報検索における有効性が示されている [2]．しかし，PLSIに基づく情報検索では，有効かつ実用的な適合性フィードバック手法の研究は見当たらない．

そこで本研究では，PLSIに基づく適合性フィードバック手法を提案する．そして，ベンチマークデータ BMIR-J2 [3] に適用することにより本手法の有効性を示す．

2 情報検索モデル

情報検索システムは，検索対象の文書集合に検索質問 q を与え，各文書と検索質問との類似度を計算することにより検索結果を得る．このプロセスには幾通りかの方法が提案されており，これらを数理的に記述するモデルがある．

2.1 ベクトル空間モデル [1]

ベクトル空間モデル (VSM) を用いた検索システムは，形態素解析処理により全文書から索引語を抽出し，この索引語を次元として，文書をベクトルで表現した索引語文書行列 A を構築する．また検索質問も，文書と同様に，索引語を次元としたベクトルで表現することができる．ベクトル空間モデルでは，索引語 $w_j(j=1,2,\ldots,J)$ を次元とした検索質問ベクトル $\boldsymbol{q}$ と各文書ベクトル $\boldsymbol{d}_i(i=1,2,\ldots,I)$ 間の類似度を余弦や内積で与えることにより，検索結果を文書を類似度の降順に並べたランキングで提示することを可能にしている．

[定義 1] 検索質問 q に対する検索質問ベクトル $\boldsymbol{q}$

$$\boldsymbol{q} = (q_1, q_2, \ldots, q_J)$$

$$q_j = \begin{cases} 0: & \text{索引語 } w_j \text{ が検索質問に含まれない} \\ 1: & \text{索引語 } w_j \text{ が検索質問に含まれる} \end{cases}$$

[定義 2] 文書 d_i に対する文書ベクトル $\boldsymbol{d}_i(i=1,2,\ldots,I)$

$$\boldsymbol{d}_i = (d_{i1}, d_{i2}, \ldots, d_{iJ})$$

$$d_{ij} = \frac{tf(w_j, d_i)}{\sum_{j=1}^{J} tf(w_j, d_i)} \times \left(\log \frac{I}{df(w_j)} + 1 \right)$$

$tf(w_j, d_i)$：文書 d_i 中の索引語 w_j の出現回数
$df(w_j)$ ：索引語 w_j が出現する文書数

[定義 3] 文書 d_i と検索質問 q 間の類似度 $sim(d_i, q)$

$$sim(d_i, q) = \frac{\sum_{j=1}^{J} q_j \cdot d_{ij}}{\sqrt{\sum_{j=1}^{J} {q_j}^2} \sqrt{\sum_{j=1}^{J} {d_{ij}}^2}} \tag{1}$$

2.2 確率モデル [4]

各文書 d_i が検索質問に適合する確率（適合確率）$P(R|d_i)$ と適合しない確率（不適合確率）$P(\bar{R}|d_i)$ の比によって，文書 d_i の検索質問に対する類似度を計算し，文書を順序付けする．ここで，R は検索質問に対する適合文書集合を，$\bar{R}$ はその否定，つまり不適合文書集合を表す．

[定義 4] 文書 d_i と検索質問 q 間の類似度 $sim(d_i, q)$

$$sim(d_i, q) = \log \frac{P(R|d_i)}{P(\bar{R}|d_i)} \tag{2}$$

2.3 潜在意味モデル

2.3.1 LSI

索引語文書行列は高次元かつスパースであるという性質を持っている．高次元スパースな行列をそのままの形で計算機上に格納するのは非効率であるため，実際に文書検索システムを構築するに際しては，何らかの方法でスパース行列を圧縮するのが望ましい．S. Deerwesterらは，意味的情報検索のモデルとしてLSI(Latent Semantic Indexing) を提案した [5]．LSIでは，索引語文書行列 A を特異値分解 (SVD) によって

$$A = U \Sigma V^T \tag{3}$$

*〒 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 Okubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 Japan. E-mail: adachi@hirasa.mgmt.waseda.ac.jp

と分解する．このうち，主成分の大きい方から K 個を用いて

$$\hat{A} = U_K \Sigma_K V_K^T \tag{4}$$

とすることにより，K 次元の潜在意味空間に圧縮することでノイズの除去を行う．これは，2 乗誤差を最小にする圧縮となっている．

しかし，LSI は索引語文書行列 A に idf 値などで ad-hoc な重み付けが必要であるなど，いくつかの問題がある．

2.3.2 PLSI

一方，T. Hofmann によって提案された PLSI (Probabilistic Latent Semantic Indexing)[2] は，LSI と同様の圧縮を確率論に基づいて行う手法である．

PLSI では，意味的な隠れ属性 $z_k (k = 1, 2, \ldots, K)$ のもとで，文書 $d_i (i = 1, 2, \ldots, I)$ と単語 $w_j (j = 1, 2, \ldots, J)$ の生起は独立であると考える．したがって，次式が成り立つ．

$$P(d, w) = \sum_k P(d_i|z_k) P(w_j|z_k) P(z_k) \tag{5}$$

ここで，文書 d_i における単語 w_j の実際の出現回数を $n(d_i, w_j)$ とすると，データの尤度

$$L = \sum_i \sum_j n(d_i, w_j) \log P(d_i, w_j) \tag{6}$$

を最大にする $P(z_k)$，$P(d_i|z_k)$，$P(w_j|z_k)$ を，以下の式を計算する EM アルゴリズムによって最尤推定する．

E-step

$$P(z_k|d_i, w_j) = \frac{P(z_k) P(d_i|z_k) P(w_j|z_k)}{\sum_{k'} P(z_{k'}) P(d_i|z_{k'}) P(w_j|z_{k'})} \tag{7}$$

M-step

$$P(w_j|z_k) = \frac{\sum_i n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{i,j'} n(d_i, w_{j'}) P(z_k|d_i, w_{j'})} \tag{8}$$

$$P(d_i|z_k) = \frac{\sum_j n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{i',j} n(d_{i'}, w_j) P(z_k|d_{i'}, w_j)} \tag{9}$$

$$P(z_k) = \frac{\sum_{i,j} n(d_i, w_j) P(z_k|d_i, w_j)}{\sum_{i,j} n(d_i, w_j)} \tag{10}$$

実際は，過学習を避けるため，Tempered EM[2] を用いて最尤推定を行っている．隠れ属性数 K は AIC 基準によって求めることが可能である [6]．

3 適合性フィードバック [1]

情報検索において，一度の検索でユーザが必要とする情報をすべて得られることはまれである．そこで，ユーザに検索結果を提示し，ユーザがその結果を見てシステムの挙動を変化させるようにシステムのパラメタを調整することが考えられる．このような技術を一般に適合性フィードバックと呼ぶ．

3.1 VSM に基づく適合性フィードバック手法

ベクトル空間モデルに基づく適合性フィードバックにおいて，検索質問ベクトルの索引語の重みを修正する手法はいくつか提案されているが，Rocchio の手法 [7] が基本となる．これはユーザが適合と判断した文書集合の重心ベクトルと不適合と判断した文書集合のベクトルの差分を新しい検索質問ベクトルとする手法で，以下の式で表現される．

$$q_{\text{new}} = q_{\text{org}} + \frac{1}{|R|} \sum_{d_i \in R} d_i - \frac{1}{|\bar{R}|} \sum_{d_i \in \bar{R}} d_i \tag{11}$$

q_{new}：適合性フィードバック後の検索質問ベクトル
q_{org}：適合性フィードバック前の検索質問ベクトル
$|S|$：集合 S の要素数

3.2 確率モデルに基づく適合性フィードバック手法

確率モデルにおいて，検索質問に対する検索結果を見て，各文書の適合・不適合を決めこれを繰り返すことにより，真の適合文書集合を見つけ出すような過程を考えている．この過程を確率的手法により定式化しているため，適合性フィードバックが必要不可欠となる．実際には，適合文書集合からランダムに取り出した文書に検索質問に含まれる索引語が出現する確率を求めるのに適合性フィードバックの情報が用いられる．

4 PLSI に基づく適合性フィードバック手法

本節では，提案手法である PLSI に基づく適合性フィードバック手法を説明する．

4.1 類似度の計算式

PLSI では，ベクトル空間モデルを確率的に次元圧縮しているために確率モデルとしても捉えることが可能である．初期検索における検索質問における文書 d_i の類似度 $sim(d_i, q)$ を次式で定義する [1][4]．

[定義 5] 初期検索における類似度 $sim(d_i, q)$

$$\begin{aligned} sim(d_i, q) &= \log P(R|d_i) \\ &= \log P(d_i|R) + c_1 \end{aligned} \tag{12}$$

ここで，c_1 は文書 d_i によらない定数でランキングには影響がない．式 (12) は文書 d_i が適合文書になる確率を類似度としていることを表している．類似度に文書 d_i が不適合文書になる確率を考慮していないのは，初期検索において，不適合文書を検索するような索引語の情報がないためである．ここで，S. E. Robertson と K. Sparck Jones は，定式化を行うにあたって，文書 d_i において個々の索引語の出現は独立で，文書が検索質問に適合する確率はその文書に出現する索引語と出現しない索引語の両方の情報を用いて計算されるべきである，という仮定を置いている [8]．PLSI では，文書の索引語の出現頻度を確率で捉えているため，各文書で全索引語が確率的に出現していると見なすことができる．したがって，その仮定の下で，式 (12) で与えた類似度は次式のように表すことができる．

$$sim(d_i, q) = \sum_{w_j \in q} \log P(d_i, w_j) + c_1 \tag{13}$$

すなわち，検索質問に対する適合文書集合からランダムに取り出した文書 d_i に，検索質問中に存在する索引語（検索語）が出現する確率 $P(d_i, w_j)$ から類似度は算出されることを意味している．

初期検索後，適合性フィードバックによって得られる適合文書中だけに存在する索引語（適合検索語）の集合を Q，不適合文書中だけに存在する索引語（不適合検索語）の集合を $\bar{Q}$ とする．適合性フィードバック後の文書

d_i の類似度を，適合確率と不適合確率の比で算出するものとして次式で定義する．

[定義 6] 適合性フィードバック後における類似度 $sim(d_i, q)$

$$sim(d_i, q) = \log \frac{P(R|d_i)}{P(\bar{R}|d_i)} = \log \frac{P(d_i|R)}{P(d_i|\bar{R})} + c_2 \quad (14)$$

式 (13) の導出と同様に，索引語が独立して出現するという仮定および文書の適合確率は適合検索語の出現確率から，不適合確率は不適合検索語から計算されるという仮定を置く．その仮定の下で，PLSI において各文書では全検索語が確率的に出現していると捉えると，式 (14) は以下のように表すことができる．

$$sim(d_i, q) = \sum_{w_j \in Q} \log P(d_i, w_j|R) - \sum_{w_j \in \bar{Q}} \log P(d_i, w_j|\bar{R}) + c_2 \quad (15)$$

すなわち，適合文書集合からランダムに取り出した文書 d_i に適合検索語が出現する確率と不適合文書集合からランダムに取り出した文書 d_i に不適合検索語が出現する確率との比により，文書 d_i の類似度は算出される．

4.2 検索システムの動作概要

PLSI に基づく適合性フィードバックの動作概要は以下の通りである．

1. (EM アルゴリズム) 尤度 L を最大にする $P(z_k)$, $P(d_i|z_k)$, $P(w_j|z_k)$ を求め，各文書 d_i と各索引語 w_j の共起確率 $P(d_i, w_j)$ を求める．
2. (初期検索および結果の提示) ユーザから入力された検索質問 q と各文書 d_i 間の類似度を算出し，類似度を降順に並べたランキングを検索結果としてユーザに提示する．
3. (適合性フィードバック) ユーザは上位文書に対して適合・不適合の適合判断を行い，検索システムにその情報を返す．
4. (EM アルゴリズム) 検索システムは，適合検索語と不適合検索語を次元として，EM アルゴリズムにより共起確率 $P(d_i, w_j|R)$, $P(d_i, w_j|\bar{R})$ を求める．
5. (再検索および結果の提示) 文書 d_i の類似度を再計算し，ランキングを提示する．
6. (検索の終了条件) 検索結果が検索要求を満たす場合は終了．満たさない場合はステップ 3 へ．

4.3 初期値依存性を利用した EM アルゴリズム

伊藤らは，EM アルゴリズムが初期値の近くにある局所解に収束するという性質を利用することで，文書自動分類において高い分類精度を実現している [9]．そこで，共起確率 $P(d_i, w_j|R)$, $P(d_i, w_j|\bar{R})$ を求めるために，EM アルゴリズムの初期値依存性を利用する．すなわち，隠れ属性数 $K = 2$ として，ひとつの隠れ属性 (z_1) に適合文書集合の概念を，もう一方の隠れ属性に (z_2) に不適合文書集合の概念を持たせるために，意図的に EM アルゴリズムの初期値を与えることを考える．

[初期値の与え方]

索引語 $w_j \in (Q + \bar{Q})$ として，隠れ属性 z_1 に適合文書集合の概念を持たせるために，初期値として確率値 $P(w_j|z_1)$ を以下のように与える．

$$P(w_j|z_1) = \frac{\sum_{d_i \in R} n(d_i, w_j)}{\sum_{w_{j'} \in Q} \sum_{d_i \in R} n(d_i, w_{j'})} \quad (16)$$

同様に，隠れ属性 z_2 に不適合文書集合の概念を持たせるために，初期値として確率値 $P(w_j|z_2)$ を以下のように与える．

$$P(w_j|z_2) = \frac{\sum_{d_i \in \bar{R}} n(d_i, w_j)}{\sum_{w_{j'} \in \bar{Q}} \sum_{d_i \in \bar{R}} n(d_i, w_{j'})} \quad (17)$$

また確率値 $P(d_i|z_k)$, $P(z_k)$ は以下のように与える．

$$\forall i, k \quad P(d_i|z_k) = 1/I \quad (18)$$

$$\forall k \quad P(z_k) = 1/K \quad (19)$$

すなわち，$P(w_j|z_1)$ は，適合検索語となる索引語に対しては適合文書集合の重心ベクトルの値を，不適合検索語となる索引語に対しては 0 を初期値として与えている．一方 $P(w_j|z_2)$ は，適合検索語となる索引語に対しては 0 を，不適合検索語となる索引語に対しては不適合文書集合の重心ベクトルを EM アルゴリズムの初期値として与えている．このように初期値を与えることは，EM アルゴリズムの収束速度を速める効果もある．

以上のように PLSI における隠れ属性に適合・不適合の概念を持たせることで，$P(d_i|R)$, $P(d_i|\bar{R})$ をそれぞれ $P(d_i|z_1)$, $P(d_i|z_2)$ として捉えることが可能となる．よって，文書 d_i の類似度は以下の式で与えられる．

$$\begin{aligned} sim(d_i, q) &= \sum_{w_j \in Q} \log P(d_i, w_j|z_1) - \sum_{w_j \in \bar{Q}} \log P(d_i, w_j|z_2) + c_2 \\ &= \sum_{w_j \in Q} \log P(d_i|z_1) P(w_j|z_1) P(z_1) - \sum_{w_j \in \bar{Q}} \log P(d_i|z_2) P(w_j|z_2) P(z_2) + c_2 \end{aligned} \quad (20)$$

確率値 $P(z_k)$, $P(d_i|z_k)$, $P(w_j|z_k)$ $(k = 1, 2)$ は，初期値を与えた EM アルゴリズムにより最尤推定することで求める．

5 シミュレーション

本節では，提案手法の有効性を示すためにシミュレーションを行い，結果を示す．

5.1 シミュレーション方法と条件

検索対象となる文書集合には，毎日新聞 CD-ROM'94 データ版 [10] を基に構築した情報検索システム評価用テストコレクションである BMIR-J2[3] を用いた．このテストコレクションは，5,080 文書から成り，あらかじめ検索課題と検索質問，また各課題に対する正解文書が与えられている．このうち検索課題は，BMIR-J2 が提供する検索課題のうち正解文書が偏らないよう考慮して 12 課題を選んだ．適合性フィードバックは一度だけ行い，初期検索で得られた上位 20 文書に対して適合性の判断を行った．検索精度の評価尺度としては適合率・再現率

から得られる平均適合率と適合率・再現率曲線を用いた．平均適合率は11点適合率の各再現率における適合率の平均であり，適合率・再現率の計算は以下の式で行う．

$$\text{適合率} = \frac{\text{正解文書のうち検索できた文書数}}{\text{検索文書数}}$$

$$\text{再現率} = \frac{\text{正解文書のうち検索できた文書数}}{\text{全正解文書数}}$$

5.2 シミュレーション結果

比較対象となる手法は，ベクトル空間モデルに基づく適合性フィードバック手法であるRocchioの手法として，各手法における検索精度の比較を示す．

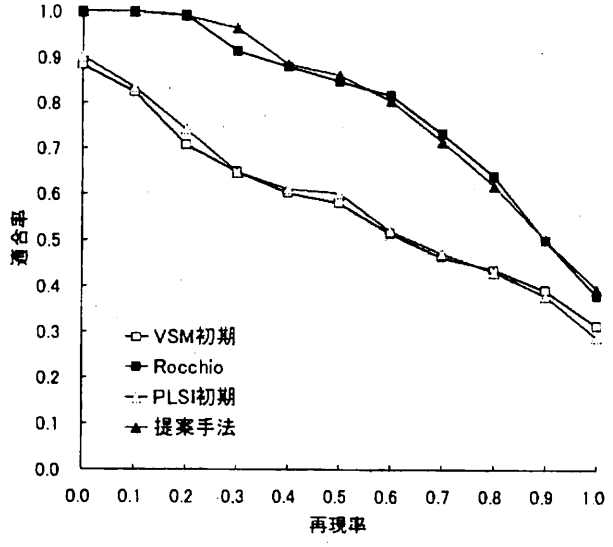

図 1: 適合率・再現率曲線

表 1: 各手法における平均適合率

	VSM 初期	Rocchio	PLSI 初期	提案手法
平均適合率	0.5784	0.7910	0.5843	0.7945

6 考察

(1) 初期検索の性能

適合率・再現率曲線から，提案手法がRocchioの手法とほぼ同等の検索精度であることが分かる．また，平均適合率で比較しても，初期検索・再検索いずれにおいても提案手法の方が，同等以上の検索精度が得られていることが分かる．初期検索において提案手法の方が検索精度の良い理由としては，ベクトル空間モデルにおいて，文書中に含まれる不必要な索引語がノイズ的な影響を及ぼし，検索精度を低下させている可能性があるためであると考えられる．PLSIでは，高次元の空間では別々に扱われていた索引語が，低次元の空間では相互に関連を持ったものとして扱われる可能性もあるため，索引語の持つ意味や概念に基づく検索を行うことができる．

(2) 適合性フィードバックの性能

適合性フィードバック後において提案手法の方が検索精度の良い理由としては，初期値を与えたEMアルゴリズムにより，文書の検索質問に対する適合確率（不適合確率）をうまく算出できたためだと考えられる．

(3) 計算時間

計算時間としては，PLSIでは，初期検索のためにEMアルゴリズムにより確率値 $P(z_k)$，$P(d_i|z_k)$，$P(w_j|z_k)$ を算出するのに掛かる時間がネックとなるが，この問題は検索前の前処理で算出しておくことで解決可能となる．適合性フィードバック後においてEMアルゴリズムで掛かる時間は，隠れ属性数は2で，対象索引語数は適合文書にのみ出現する索引語と不適合文書にのみ出現する索引語の和であるから，そこまで大きな次元数ではないため計算時間は大きく掛からない．したがって，対話的な情報検索が可能であると考えられる．

今回，提案手法において，ベクトル空間モデルに基づく適合性フィードバック手法であるRocchioの手法とほぼ同等の検索精度が得られた．しかし，ベクトル空間モデルでは高次元スパースな索引語文書行列をそのままの形で計算機上に格納するのに対し，PLSIでは索引語文書行列を次元圧縮しているため，計算機のメモリや検索時間による制約を受けにくいという利点がある．以上のことから，提案手法が有効な手法であると言える．

7 まとめと今後の課題

本研究では，PLSIに基づく適合性フィードバック手法を提案し，シミュレーションによってその有効性を確認した．

今回は文書と検索質問との類似度を確率モデルとして扱ったが，確率ベクトルに対してもベクトル空間モデルのように余弦で算出する方法やKL情報量を用いる方法などが存在する．今後は，他の類似度との比較を行い，より高精度なPLSIに基づく適合性フィードバック手法を検討していきたい．

8 謝辞

著者の一人である足立は，本研究を行うにあたり，数多くのご助言，ご支援を賜りました早稲田大学平澤研究室の各氏に感謝いたします．

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”，共立出版株式会社，(2002).
[2] Hofmann, T. :“Probabilistic Latent Semantic Indexing”, *Proc. of SIGIR'99, ACM Press*, pp.50-57, (1999).
[3] 情報処理学会データベースシステム研究会:“BMIR-J2 テストコレクション”，新情報処理開発機構，(1998).
[4] 徳永健伸:“情報検索と言語処理”，東京大学出版会，(1999).
[5] Deerwester, S. , Dumais, S. T. , Furnas, G. W. , Landauer, T. K. and Harshman, R. : “Indexing by Latent Semantic Analysis”, *J. of the Society for Information Science*, 41, pp.391-407, (1990).
[6] 赤池弘次:“情報量基準 AIC とは何か—その意味と将来への展望”，数理科学，No.153, pp5-11, (1976).
[7] Rocchio, J. :“Relevance Feedback in Information Retrieval”, *The SMART Retrieval System Experiments in Automatic Docment Processing, Prentice Hall Inc*, (1971).
[8] Robertson, S. E. & Sparck Jones, K. :“Relevance weighting of search terms”. *Journal of the American Society for Information Science*, 27(3), pp.129-146, (1976).
[9] 伊藤潤，石田崇，後藤正幸，酒井哲也，平澤茂一:“PLSI を利用した文書からの知識発見”，情報科学技術フォーラム，Vol.2, D-039, (2003).
[10] 毎日新聞社:“CD-毎日新聞'94 データ集”，日外アソシエーツ，(1994).