自 然 言 語 処 理　135－3
（2000．1．27）

教師あり学習によるベクトル空間モデルの精度改善

Tai Xiaoying　　北 研二

徳島大学 工学部
〒770-8506　徳島市南常三島町 2-1
{xytai,kita}@is.tokushima-u.ac.jp

概要

本稿では、少数の検索質問に対しどの文書が正解であるかという情報を用いて、ベクトル空間モデルに基づく情報検索システムの精度を改善する方法を提案する。また、LSI(Latent Semantic Indexing) に基づく情報検索システムに提案した方法を組み入れ、2 つのテスト・コレクション (Medline と Cranfield) を用いて実験を行った。LSI と比べ、学習データに対し、それぞれ 4.03%と 26.79%の検索精度 (平均精度) の向上が得られ、テストデータに対しては、0.01%と 5.59%の検索精度の向上が得られた。

Improvement of Vector Space Model based on Supervised Learning

Xiaoying Tai　　Kenji Kita

Faculty of Engineering, Tokushima University
2-1, Minami-josanjima, Tokushima 770-8506, Japan
{xytai,kita}@is.tokushima-u.ac.jp

Abstract

In this paper, a method is proposed to improve retrieval performance based on the vector space model (VSM) by using the information about that which documents are right answers to the small number of questions. In addition, an approach is incorporated which is proposed for the information retrieval system based on LSI(Latent Semantic Indexing) and experiments were carried out using two test collections (Medline and Cranfield). The rises of retrieval precision compared with LSI were 4.03% and 26.79% for two training data, and 0.01% and 5.59% for test data, respectively.

1 はじめに

ベクトル空間モデルは情報検索の代表的なモデルである。ベクトル空間モデルでは、文書あるいは検索質問を索引語の重みベクトルで表現する。また、検索質問と各文書間の適合度をベクトル間の類似度により計算する。ベクトル空間モデルの特徴として、適合度に基づき検索結果に対する順位付けができること、および検索質問に含まれる索引語に重みを与えることができることなどを挙げることができる。

ベクトル空間モデルにおいて、ユーザからのフィードバック情報を利用する方法として、適合性フィードバック (relevance feedback) と呼ばれる技術がある。適合性フィードバックとは、システムがユーザに検索した結果を提示し、ユーザがその結果の適合性を判断して、システムの動きを変化させるようにシステムのパラメータを調整することである。適合性フィードバックが調整する対象としては、検索モデル、文書、検索質問がある。通常は、検索質問の重みを修正するか、検索質問の拡張を行う。しかし、単に検索質問を修正するだけでは、システムに対して長期的な効果を残すことができない。

ユーザからフィードバック情報を利用する他の方法としては、ユーザレンズ (User Lens) がある [1, 4]。ユーザレンズとは、"Guttman's Point Alienation Statistic" と呼ばれる尺度に基づき、ベクトルの重みを修正する線形変換である。この方法では、システムが出力したランキング結果をもとに、文書と (あるいは) 質問の重みを改めて計算することにより、より適合するランキング結果を得ることができる。したがって、質問だけではなく、文書の重みを修正することにより、ユーザからのフィードバック情報を長期的に残すことができるという特徴がある。

本稿では、少数の検索質問に対してどの文書が正解であるかという情報を用いて、ベクトル空間モデルに基づく情報検索システムの精度を改善する方法を提案する。検索精度を高めるため、ユーザからの適合性の判定結果 (どの文書が正解であるか) だけではなく、文書間距離などの情報もモデルに組み入れ、線形変換によりベクトル空間検索モデルを構築した。2 つのテストコレクションを用いて、評価実験を行った結果、学習データに対してもテストデータに対しても検索精度の向上が得られた。また、本稿で提案する手法は、システムにユーザからのフィードバック情報に線形変換を行い、長期的にシステムに反映させる点ではユーザレンズと類似しているが、モデルを調整する方法がユーザレンズとは異なっている。

2 線形変換による検索モデルの構築

2.1 ベクトル空間モデルに基づく検索

ベクトル空間モデルでは、文書 $d_j (1 \leq j \leq n)$ を索引語の重みベクトル

$$d_j = (w_{1j}, ..., w_{tj})^T \tag{1}$$

で表現する。ここで、t は索引語の総数であり、w_{kj} は文書 d_j における k 番目の索引語の重みである。n 個 の文書ベクトルは、次の単語・文書行列 (term-document matrix) で表現される。

$$D = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{t1} & w_{t2} & \cdots & w_{tn} \end{bmatrix}$$

検索質問も、文書と同様に索引語の重みベクトルで表現する。各質問

$$q = (q_1, ..., q_t)^T \tag{2}$$

が与えられたとき、各文書 d_j との類似度 $sim(d, q)$ は次式

$$sim(d, q) = \sum_{k=1}^{t} (w_{kj} \times q_k) \tag{3}$$

で計算することができる。検索質問ベクトルの集合 Q に対する検索結果は、類似度行列

$$\Re = D^T Q \tag{4}$$

によって表される。

2.2 検索モデルに関する線形変換

検索質問集合の行列 Q が与えられたとき、各質問に対しどの文書が正解であるかという情報を正解行列 R によって表現する。正解行列 R の要素 r_{ij} には、質問 q_i に対して、文書 d_j が正解であるときに 1 を、正解でないときには 0 を与える。

ここで線形変換 $\mathbf{L}$ により、

$$R = \mathbf{L}[Q] \tag{5}$$

は、検索質問集合行列 Q と正解行列 R を関連付けることを考える。このとき、検索モデルは

$$R = D^T X Q \tag{6}$$

によって表すことができる。X は、$\mathbf{L}$ の線形変換により導入した行列である。

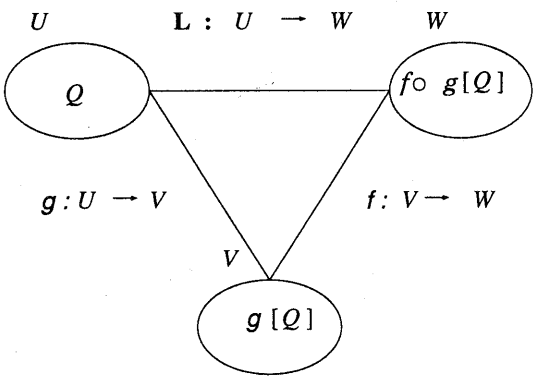

図 1: 検索モデルに関する線形変換

検索モデル (6) は、X の導入によって線形空間 U(各質問ベクトル $q \in U$) の基底が線形空間 W (各正解ベクトル $r \in W$) の基底に変換されることを示している。いま、図 1 に示されるように、g を線形空間 U から V への線形変換とし、f を線形空間 V から W への線形変換とする。$\mathbf{L} = f \circ g$ は合成線形変換で、線形空間 U から W への線形変換である。ここで、

$$M = XQ = g[Q] \tag{7}$$

とする。U, V, W は、$dim(U) = dim(V) = t, dim(W) = n$ のベクトル空間とし、それぞれ $(e_1, ..., e_t)$, $(e_1^{'}, ..., e_t^{'}), (e_1^{''}, ..., e_n^{''})$ の基底を持っているとする。g の表現行列が X であり、f の表現行列が D である。U, V, W の基底を定めたとき、U から V への線形変換と g の表現行列 X とは 1 対 1 に対応し、V から W への線形変換と f の表現行列 D とも 1 対 1 に対応する。したがって、$\mathbf{L} = f \circ g$ の変換により線形空間 U における各質問ベクトル q が与えられたとき、どの文書が正解であるかという情報を表している正解ベクトル空間 W に変換されるわけである。次式は、この変換過程を示す。

$$\begin{aligned}
\mathbf{L}[Q] &= f \circ g[Q] \\
&= f \circ g\left[\sum_{k=1}^{t} Q_k e_k\right] \\
&= f\left[\sum_{k=1}^{t} Q_k \Big(g(e_k)\Big)\right] \\
&= f\left[\sum_{j=1}^{t} \Big(\sum_{k=1}^{t} x_{jk} Q_k\Big) e_j^{'}\right] \\
&= f\left[\sum_{j=1}^{t} M_j e_j^{'}\right] \\
&= \sum_{j=1}^{t} M_j \Big(f(e_j^{'})\Big) \\
&= \sum_{i=1}^{n} \Big(\sum_{j=1}^{t} d_{ik} M_j\Big) e_i^{''} \\
&= \sum_{i=1}^{n} R_i e_i^{''}
\end{aligned} \tag{8}$$

2.3 X 行列の計算

行列 X は、検索質問と正解を関連付ける行列である。まず、式 (6) および式 (7) より

$$R = D^T M \tag{9}$$

となる。行列 X を求めるためには、式 (9) を満たす M を求め、次に、式 (7) を満たす X を求める。しかし、一般に式 (9) の連列一次方程式は解が存在しない。このため、次の最小 2 乗問題を解くことにより、最適解 M^* を求める。

$$M^* = \underset{M}{argmin} \, \|R - D^T M\|_F^2 \tag{10}$$

上式において、$\|\cdot\|_F$ はフロベニウス・ノルムである。式 (10) を解くため、まず、QR 分解を用いて、D^T を $D^T = \mathrm{Q}R$ に分解する。M^* は、

$$\begin{aligned} y &= \mathrm{Q}^T R \\ M^* &= \mathrm{R}^{-1} y \end{aligned} \tag{11}$$

で計算される。この解 M^* は、$R = D^T M$ に対する最小2乗の意味で最適な解である。M^* を求めてから、次に $M^* = M = XQ$ とし、次の最小2乗問題を解くことにより、最適解 X^* を求める。

$$X^* = \underset{X}{argmin} \|M - XQ\|_F^2 \tag{12}$$

2.4 文書間距離の導入

モデル (6) を用いて、Medline コレクションに対して予備的な実験を行ったが、学習データに対しては全ての再現率において 100% 近い精度に達したが、テストデータに対しては逆に検索精度を落す結果となった。テストデータの精度を上げるためには、さらにモデルに何らかの情報を組み入れる必要がある。

一方、文書と文書の間の相関 (文章間距離) は自己相関行列 $D^T D$ によって表される。我々は、検索質問に対する正解情報に加え、上の文書間の相関情報を、検索モデルに組み入れる方法について検討した。

まず、2つの行列 X, Y に対し、X と Y の要素を順に並べてできる行列 Z を X と Y の合弁と呼び、

$$Z = X \circ Y \tag{13}$$

と表すことにする。例えば、

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

$$Y = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$$

であるとき、

$$Z = \begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{bmatrix}$$

となる。

検索モデル (6) に文書間相関行列 $D^T D$ を組み入れるため、我々は、式 (4) の $\Re = D^T Q$ に対して行列 $D^T D$ を組み入れ、その結果をモデル (6) に適応した。まず、式 (4) を次のように

$$\Re \circ (D^T D) = (D^T Q) \circ (D^T D) = D^T (Q \circ D) \tag{14}$$

拡張することができるので、さらに行列 X を導入することにより、検索モデル (6) を次のように拡張する。

$$R \circ (D^T D) = D^T X (Q \circ D) \tag{15}$$

2.5 LSI による概念空間への変換

上に述べた検索モデルの次元を削減するために、我々は さらに LSI(Latent Semantic Indexing)[2] を用いた。LSI は、高次元行列を特異値分解 (Singular Value Decomposition SVD) により縮退し、低次元のベクトル空間、つまり索引語と文書の潜在的な意味構造を表した概念空間に変換する技術である。LSI を用いることにより、メモリの節約と検索時間を減少することができるだけではなく、情報検索の性能を改善することができる。

単語・文書行列 D は、SVD を用いること により、

$$D = U_{(t,r)} \Sigma_{(r,r)} V^T_{(r,n)} \ (r = rank(D)) \tag{16}$$

のように展開される。ここで、U と V は直交行列 ($U^T U = V^T V = I$) であり、また、Σ は対角行列であり、対角要素には特異値が降順に並んでいる。LSI においては、$(k < r)$ の大きな特異値だけを用いて、行列 D に対する近似行列 D_k を再構成する。

$$D_k = U_k \Sigma_k V_k^T \ (k < rank(D)) \tag{17}$$

D_k は SVD の結果からノイズに相当する特異値を削減し、概念空間において再び合成したものであるため、D と比べ索引語と文書の間の関連をより的確に表現している。

最終的に、我々は次の検索モデルを用いた。

$$R_k \circ (D_k^T D_k) = D_k^T X_k (Q_k \circ D_k) \tag{18}$$

3 評価実験

3.1 実験概要

検索モデル (18) に対する評価実験を Medline コレクションと Cranfield コレクションを用いて行っ

た。Medline コレクションと Cranfield コレクションは、それぞれ医学と航空学の抄録からなる文書集合である。これらのデータをまず、以下のように処理をした (表 (1) 参照)。

表 1: 評価実験に用いたデータの規模

	Medline	Cranfield
文書数 n	1033	1400
検索質問数 l	30	225
抽出された単語数 m	4329	4991

2 つのコレクションから停止語 (439 個) 処理と接辞処理アルゴリズム [3] を用いて、索引語を抽出する。n 個の文書ベクトル $d_j = (d_1, ..., d_n)$ を作る。文書ベクトル d_j の第 i 番目 の要素は

$$d_{ij} = L_{ij} G_i \tag{19}$$

である。L_{ij} は文書 d_j の第 i 番目の単語の局所的重みであり、G_i は 第 i 番目の単語の大域的重みである。ここで、L_{ij} は

$$L_{ij} = 1 + \log f_{ij} \tag{20}$$

であり、G_i は

$$G_i = 1 + \sum_{j=1}^{n} \frac{\frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}}{\log n} \tag{21}$$

である。この式において、f_{ij} は単語 i が文書 j に出現する頻度で、F_i は単語 i が文書集合全体にわたり出現する頻度である。

Medline の検索質問集合行列 Q と正解集合行列 R を Q_1, R_1 (各 20 個) と Q_2, R_2 (各 10 個) に分割し、Cranfield の検索質問集合行列 Q と正解集合行列 R を Q_1, R_1(各 169 個) と Q_2, R_2 (各 56 個) に分割した。どちらでも Q_1, R_1は学習データで、Q_2, R_2 はテストデータである。トレーニングするとき、Medline と Cranfield コレクションの単語 ・文書行列 D、質問行列 Q と正解行列 R は 2章で述べたようにそれぞれ LSI によって k (= 100) 次元の概念空間に変換した。

3.2 実験結果

実験において、式 (18) を用い、X_k を求める。最適な X_k を求めるため、正解行列 R_k の重みをいろいろ変化させ、トレーニングした。Medline コレクションに対しては、その重みを 1.0 とし、Cranfield コレクションに対しては、その重みを 10.0 とし、X_k を求めた。求めた X_k を用い、学習データ (Q_1) とテストデータ (Q_2) の実験結果を検証する。

表 2に各重みについての検索した平均精度および LSI の検索平均精度との比較を示す。

表 2: LSI と検索結果の平均精度比較

		Medline		Cranfield	
	データ	重み	平均精度	重み	平均精度
LSI	学習	-	0.6747	-	0.4001
	テスト	-	0.6927	-	0.4434
提案した検索モデル	学習	0.1	0.6806	2.0	0.4280
		0.15	0.6822	3.0	0.4413
		0.2	0.6833	4.0	0.4564
		0.5	0.6915	5.0	0.4650
		1.0	0.7019	10.0	0.5073
	テスト	0.1	0.6930	2.0	0.4463
		0.15	0.6933	3.0	0.4591
		0.2	0.6933	4.0	0.4591
		0.5	0.6931	5.0	0.4610
		1.0	0.6928	10.0	0.4682

図 2と図 3 に Medline コレクションを用いた実験結果を、図 4と図 5に Cranfield コレクションを用いた実験結果を示す。LSI と比べ、Cranfield の学習データについても、テストデータについても精度改善が見られた。Medline の学習データについて精度改善が見られたが、テストデータについては精度改善があまり見られなかった。Medline データと Cranfield データの学習データに対し、それぞれ 4.03%と 26.79%の検索精度の向上が得られ、テストデータに対し、それぞれ 0.01%と 5.59%の検索精度の向上が得られた。

4 おわりに

我々は、線形変換により、少数の検索質問に対してどの文書が正解であるかという情報や文書間距離情報をモデルに組み入れ、教師あり学習によるベクトル空間モデルを構築し、精度改善を試みた。その結果、精度改善の向上が認められ、我々の提案したモデルの有効性を確認することができた。さらに、このモデルは、検索モデルに組み入れた情報を長期的に保存するという利点がある。

今後が検索モデルのより一層の性能改善を試みたい。

参考文献

[1] Bartell, B. T., Cottrell, G. W. and Belew, R. K.: "Optimizing Parameters in a Ranked Retrieval System Using Multi-Query Relevance Feedback", *Proceedings of the Symposium on Document Analysis and Information Retrieval*, Las Vegas, 1994.

[2] Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., Harshman, R.: "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391-407, 1990.

[3] Frakes, W. B. and Baeza-Yates, R.: *"Information Retrieval: Data Structures and Algorithms"*, Prentice Hall, 1992.

[4] Vogt, C. C., Cottrell, G. W., Belew, R. K. and Bartell, B. T.: "User Lenses − Achieving 100% Precision on Frequently Asked Questions", *Proceedings of User Modeling'99*, Banff, 1999.

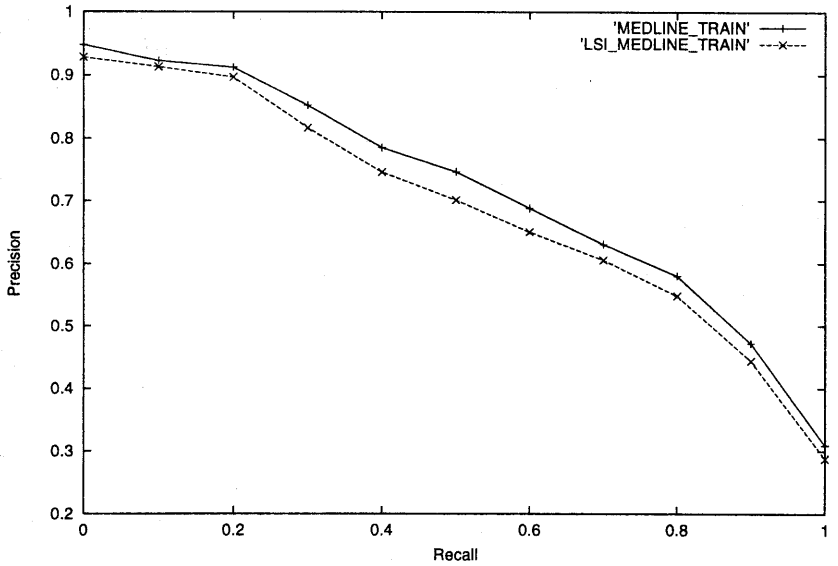

図 2: Medline コレクション (学習データ) に対する実験結果

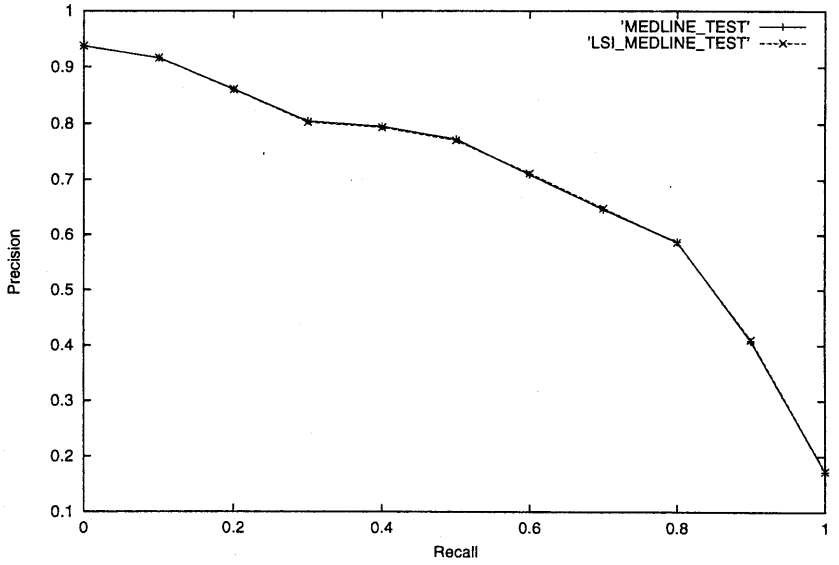

図 3: Medline コレクション (テストデータ) に対する実験結果

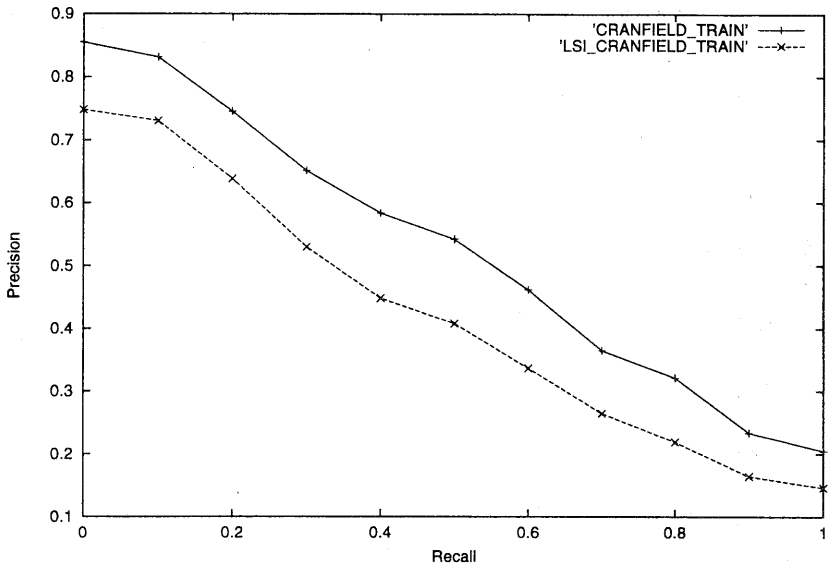

図 4: Cranfield コレクション (学習データ) に対する実験結果

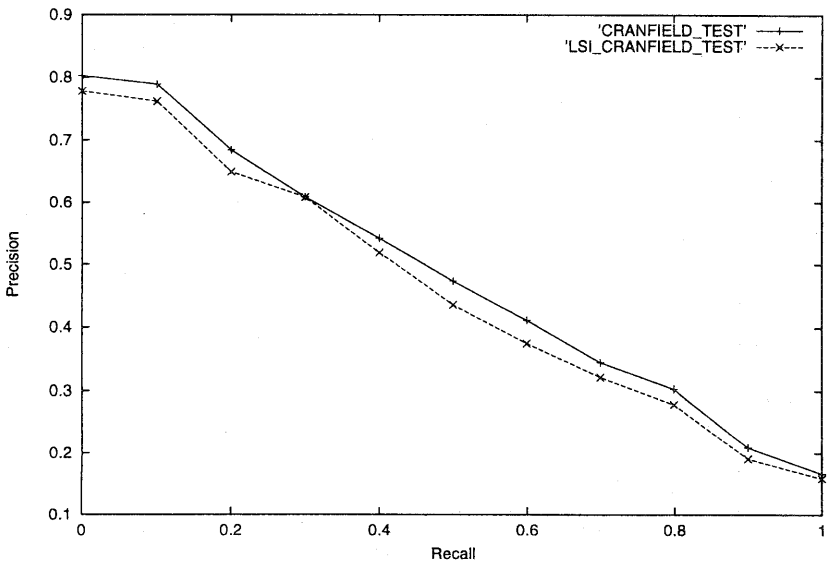

図 5: Cranfield コレクション (テストデータ) に対する実験結果