自 然 言 語 処 理 148－4
（2002．3．4）

統計的潜在的意味空間の抽出

川前 徳章† 青木 輝勝† 安田 浩†

†東京大学 先端科学技術研究センター
〒153-8904 東京都目黒区駒場 4-6-1

E-mail: †{kawamae,aoki,yasuda}@mpeg.rcast.u-tokyo.ac.jp

あらまし 本研究は統計的潜在的意味のインデキシング(SLSI)という新しい数学的アプローチを提案する。提案手法は文書だけでなく、文書内に出現した単語も同時に潜在的意味空間に配置する事ができ、その空間においてインデキシングを行える。これは潜在的意味が文書に出現した単語よりも文書をインデキシングできるためである。特異値分解に基づいた LSI やそれを発展させた PLSI と SLSI の相違点はずっと意味のあり因子分析と情報理論に基づいた堅固な統計モデルをもっていることである。それゆえ SLSI は LSI や PLSI で未解決だったいくつかの問題点を解消する事ができた。テストコレクションについてこの実験を行った結果、SLSI は LSI や PLSI よりも精度が良かった。加えてエントロピーに基づいた単語の重み付けを提案し、これを利用した結果、我々は事前に重要な単語を判断し、その結果文書中に出現した全単語から最小限必要な単語を選択する事ができる。従って、この手法は計算コストの減少を実現する事を可能とした。

キーワード 情報検索 単語選択 概念検索 特異値分解 因子分析 潜在的意味

Extraction of Statistical Latent Semantic Space

Noriaki KAWAMAE†, Terumasa AOKI†, and Hiroshi YASUDA†

†Research Center for Advanced Research and Technology The University of Tokyo
4-6-1, Komaba, Meguroku, Tokyo, 153-8904, JAPAN

E-mail: †{kawamae,aoki,yasuda}@mpeg.rcast.u-tokyo.ac.jp

Abstract The main goal of this paper is to propose Statistical Latent Semantic Indexing (SLSI) that is a novel statistical approach to simultaneously map documents and terms into a latent semantic space. This is because latent semantics of the documents fits to categorize the documents than indexing terms in the documents. In contrast to Latent Semantic Indexing (LSI) based on Singular Value Decomposition (SVD) and Probabilistic Latent Semantic Indexing (PLSI), SLSI has a more meaningful and solid statistical model that is based on a factor analysis and information theory. Therefore, this model can solve the remained critical problems in LSI and PLSI. Experimental results with a number of a test collection show that SLSI is much better than LSI and PLSI in viewpoints of retrieval. Moreover, we propose a new term weighting method based on entropy. By this method we can judge which terms are important, and can extract only minimum essential terms from them. As a result, this method makes it possible to reduce calculation cost.

Key words Information Retrievalnformation Retrieval, Words Selection, Conceptual Search, Singular Value Decomposition, Factor Analysis, Latent Semantic

1 はじめに

近年,インターネットやパソコンの普及に伴い，我々がオンラインで入手できる電子化された情報の量が増大し、それに比例して検索時間も増大していくことが予想される。その結果，我々が効率的に必要な情報を入手するために，検索システムの重要性はますます高まっていくと考えられる。

しかし，www を対象にした現在の検索はユーザにとって情報入手は効率的でない。その原因は，本来，www を対象とする検索システムがデータベースを対象とする検索システムとデータ構造，ユーザ層，検索の目的が異なるにも関わらず，現状はデータベースを対象とした検索システムの転用に過ぎないのでデータの検索に留まっていることにある.

ここでデータと情報の違いは、データは全ての人にとって同じように解釈されるのに対し、情報は人の感性や意味付けが伴うので、人によって異なった概念を持つ。従ってデータは表現そのものが変わってしまうと意味が変わってしまうが、情報は表現が異なっても同じ意味を表す場合がある。

そこでユーザの情報検索を効率化するために，検索システムは検索モデルにおいて従来の文書のインデキシングを従来の単語単位のインデックスから別のインデックスで行うことが必要になる．このインデキシングを行う空間を潜在的意味空間と呼び LSI[2],[3],[4]，PLSI[5]などが提案されてきたが、いくつかの問題を残している。

本研究はコードモデルに基づいた新しいインデキシング手法として SLSI(Statistical Latent Semantic Indexing)を提案する．SLSI は潜在的意味空間の抽出に因子分析を導入する．(P)LSI では出現した潜在的意味空間が単語の要約であったのに対して，因子分析では処理の前段階で潜在的意味を仮定したコードモデルを設定して行うことができるためである．更に因子分析の導入にあたって潜在的意味の抽出に有効な単語の重み付けを提案し，因子得点の計算方法と統計モデル選択指標の提案により従来の因子分析の問題点も解決する．提案手法の成果を次に示す.

■因子分析による潜在的意味空間の抽出
既存手法より単語のゆらぎや表記の違いに頑健な内容レベルでの文書の検索を実現した.

■統計モデル選択指標による統計モデルの検証
因子分析における最適な統計モデルは仮定した統計モデルの中で対象となるデータに一番当てはまるものである．この統計モデルの当てはまりの良さを客観的に評価する指標を MDL 基準から導出し，最適な潜在的意味空間の検証を実現した.

■単語の選択
本稿で提案した単語の重み付けは，文書分類に有効なだけでなく，その分類に必要な単語を文書集合に出現した全単語から抽出し，結果，計算量の逓減を実現した.

2 文書のインデキシングに関する従来研究とその問題点

検索システムにおける問題点を解決するために検索モデルを改善することが必要になる．具体的には検索対象となる文書のインデキシングを従来の単語単位のインデックスから別のインデックスで行うものである．このアプローチの代表的なものに LSI(Latent Semantic Indexing)がある。LSI はベクトル空間モデルの空間を基にしている。ベクトル空間モデルは文書間の関係を幾何学的に考察することを可能にするモデルである．このモデルによって，文書は，文書に出現する単語を軸とする空間において，それぞれが含んでいる単語の重みによって配置される．文書がこの空間に配置されることで，文書間の類似度はこの空間における位置関係ことで測定することができるようになる．これが従来の単語単位のインデキシングであるが、LSI は特異値分解 SVD(singular value decompostion)に基づいて，元の単語の軸よりも少数の軸で構成された空間に文書を配置するものである．従って文書は LSI によって単語単位からこの少数の軸によってインデキシングされるようになる．この LSI における軸は，ベクトル空間モデルにおける単語の軸を合成した軸であり，単語の潜在的意味の軸であると考えられているが、次のような問題点がある.

■SVD による潜在的意味空間の導出
SVD は与えられた行列のランク数で最小二乗法によって最小となる行列の再構成を行う手法である。従って LSI は SVD に基づいているので顕在変数，つまり出現した単語の合成になっている．従って，与えられた文書に出現した単語の要約に

過ぎず，潜在的意味を求めるアプローチになっていない。

■仮説の検証
LSI は仮説となるモデルの存在を前提としておらず観測データの要約に過ぎない。従って仮説検証的なアプローチになっていないために，その抽出された潜在意味空間の最適性は保証されない.

3 潜在的意味空間における文書のインデキシング

本論文では文書の検索を、文書の持つ内容の類似性により行うために潜在的意味空間を抽出する新しい手法として SLSI を提案する。なぜなら潜在的意味は文書に出現した単語よりも文書の主題や概念に関連している。従って、文書を潜在的意味によってインデキシングすることで、文書の検索や文書間の類似性を内容の類似性に基づいて行えることが実現される。SLSI は従来の手法とは利用する統計モデルと目的関数が異なる。本章ではこの潜在的意味を抽出する統計モデルとしてコードモデルを提案する。コードモデルは潜在的意味と単語及び、文書の関連性を定義したモデルであり，モデルを統計的に解く手法に因子分析を利用する。その結果、コードモデルによって文書が潜在的意味によってインデキシングされるようになるだけでなく，従来手法で未解決だった諸問題を解決することができた。

3.1 単語・文書行列とベクトル空間モデル

ベクトル空間モデルは単語を軸とする空間で文書をベクトル表現するモデルである。文書は形態素解析によって単語を要素とするベクトルで表現でき、対応する単語の重みを座標とすることでベクトル空間に文書を配置することができる。その結果、このモデルによって文書間類似度をベクトルの類似度によって幾何学的に定義することができる。ベクトル空間は次のような行列形式で表現できる。ここで行列 A を単語・文書行列、d を単語を要素とする文書ベクトル、t を文書を要素とする単語ベクトルとして表す。

$$A=\begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix} \equiv (d_1 \ \cdots \ d_n) \equiv \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix}$$

ベクトル空間を表現する単語・文書行列は次の三要素から構成されている。

(1)局所的重み付け(Local Weighting)
(2)大局的重み付け(Global Weighting)
(3)類似性の指標

(P)LSI だけでなく提案手法の SLSI もベクトル空間モデルに基礎を置いている。その違いは最終的に文書を配置する空間を構成する軸が単語でなく潜在的意味で構成されている点である。従って、文書が単語を軸とするベクトル空間から潜在的意味を軸とする潜在的意味空間に配置される。潜在的意味は単語よりも文書の概念に関連が高い。その結果、潜在的意味空間における文書の類似関係は単語を軸とするベクトル空間より文書の内容の類似性と密接な関連を持つ。次に文書の集合から潜在的意味の因子を推定するために、以下で単語・文書行列とそれから潜在的意味を抽出するモデルとしてコードモデルの定式化を行う。

局所的重み付けは、各文書内において単語の重み付けを行う。従って、同じ単語であっても、出現する文書が異なれば局所的重み付けによる値も違ってくる。本稿は文書内における単語の出現頻度は文書の内容を反映するものと考えて次の重み付けを提案する。

$$L_{ij} = P_{ij} * \log(1+P_{ij})$$

大局的重み付けは文書集合全体に渡って単語の重み付けを行う。従って、どの文書においても同じ単語であれば、同じ値をとる。本稿は文書集合全体における単語の各文書における出現頻度の違いが文書の内容を反映するものと考えて次の重み付けを提案する。

$$G_j = 1 - \frac{1}{\log N} \sum_{i=1}^{N} p_{ij} \log p_{ij}$$

3.2 SLSIによる潜在的意味空間の抽出

3.2.1 コードモデル

本稿では潜在的意味空間と文書及び文書内の単語の関係をコードモデルで定義する。コードモデルは文書に出現する単語は潜在的意味によって決定されると考えた。この考えは情報理論におけるデータと情報源の関係に似ていると考えられる。そこで本稿ではこのモデルをコードモデルと呼ぶ。本稿は文書内の単語はある潜在的意味を記号化したものと考える。そこでコードモデルにおいて潜在的意味因子 l_k と単語 w_j の関係を次のように定義する

$P(w_j|l_k)$:潜在的意味因子 l_k において単語 w_j が出現する確率

ここで$\sum_{j=1}^{m} P(w_j|l_k)=1$とする。また単語 w_j は特定の潜在的意味因子 l_k だけでなく、他の潜在的意味因子 l_kからも出現する事が可能である。

次にある文書 d_i は潜在的意味因子 l_k に対して確率的に属すると考える。そこでコードモデルにおいて潜在的意味因子 l_k と文書 d_i の関係を次のように定義する。

$P(l_k|d_i)$:文書 d_i が潜在的意味因子 l_k に属する確率

ここで$\sum_{i=1}^{n} P(l_k|d_i)=1$とする。また文書 d_i は特定の潜在的意味因子 l_k だけでなく、他の潜在的意味因子 l_kに対しても属する事ができる。

これらの確率を結合する事で、コードモデルにおいて、文書 d_i における単語 w_j の出現確率は次のように定式化できる。

$$P(w_j|d_i)=\sum_{k=1}^{l} P(w_j|l_k)P(l_k|d_i)+P_\varepsilon(w_j)P_\varepsilon(d_i)$$

実際の文書に出現する単語の確率は潜在的意味因子だけでは説明できず、単語 w_j 及び文書 d_i に固有な出現確率が存在するとして $P_\varepsilon(d_i)$と $P_\varepsilon(w_j)$ の確率を加えて以上のようにコードモデルを定式化する。従って、このコードモデルの目的関数はこの値を少ない潜在的因子の数 l で $P_\varepsilon(d_i)$と $P_\varepsilon(w_j)$の値を小さくする事である。また、ここで個々の単語 w_j と文書 d_i は同じに複数の潜在的意味因子に対して関係を持ち、潜在的意味因子間には独立という制約が無いと言う点に注意する。

3.2.2 特異値分解

3.2.2.1 LSI

LSI は特異値分解に基づく手法である。単語・文書行列 A が与えられた場合、この行列 A は特異値分解によって次のように表す事が出来る。

$$A=U\Sigma V^T$$

ここで行列 U は行列 AA^T の固有ベクトルの行列であり、行列 V も同様に行列 AA^T.の固有ベクトルの行列であり，$U^TU=I$, $V^TV=I$ (I は単位行列)の関係を満たす。行列∑は固有値の直交行列である。r は行列 A のランク(n, m)の小さい方の数とすると，U は n*r 行列, V は m*r 行列. ∑は r*r 行列となる。ここで∑の大きい順に k 個だけを使って再構成した行列 A_k は次のように求められる。

$$A=U\Sigma V^T \fallingdotseq U_k\Sigma_k V_k^T=A_k$$

U_k および V_k は最初の k 個の左あるいは右から成る行列である。A_k はランク k の行列で A の最小二乗法において最良の行列である。

3.2.2.2 PLSI

PLSI は LSI に確率分布の解釈を与えたものである。行列 A は文書 d における単語 w の出現確率として、次のように表せる事を主張している。

$$P(w|d)=P(w|z)\,P(z)\,P(d|z)$$

この式において P(w|z) は行列 U, P(z)は行列Σ, P(d|z)は行列 V の要素と解釈を与えている。LSI

も、それに確率分布の解釈を与えたPLSIにしても与えられた単語・文書行列Aを最小二乗法において最良の行列A_kを求めるもので，潜在的意味因子を仮定していない。結果として得られたzを潜在的意味因子として扱っている。特異値分解に基礎を置いた従来の研究は提案したモデルに比べて次の二点で欠けている。

(1)仮定の不在：文書の単語の出現に関して何ら仮定を置いていないので、その出現の原因となる概念の推測が出来ない。

(2)モデル選択指標の不在：従来は次元縮小の基準は恣意的であった。特異値分解による近似は最小二乗誤差を保証するが、問題はこのk個の決定である。数が多ければ、SVDをする意味がなくなり、少なければ元の単語・文書行列Aとの誤差が大きくなる。

3.2.3 因子分析モデル

統計の一手法に因子分析がある。因子分析は観測されたデータから、それらの原因となる少数の因子を発見する手法である。因子分析を用いて単語・文書と潜在的意味因子を定式化すると次のようになる。

$$a_{ij}=w_{i1}c_{1j}+w_{i2}c_{1j}+\cdots+w_{in}c_{nj}+u_iv_i$$

a_{ij}:文書d_iにおける単語w_jの観測値
先に挙げた単語・文書行列Aの重みに対応
c_{mj}:因子得点。単語w_iにおける潜在的意味因子c_mの得点
w_{im}:因子負荷量。単語w_iと潜在的意味因子c_mの相関
v_i:独自因子得点。文書d_iに固有な得点
u_i:独自因子負荷量。文書d_iと独自因子得点v_iの相関
$m \leqq j$: 潜在的意味因子の個数は単語の総数よりも小さい

以上より、単語・文書行列Aは次の形式で表現できる

$$A=WC+VU$$

W:共通因子パターン行列、(i×m)型行列。
C:共通因子行列、(m×j)型行列。
U:独自因子パターン行列、(i×i)型行列。対角成分のi番目が文書d_iの独自因子負荷量、他の成分は0。
V:独自因子得点行列、(i×j)型行列。

以上より、コードモデルを解くには、特異値分解より因子分析が適当な事が分かる。

3.2.4 幾何学的解釈

(P)LSIと因子分析を導入したSLSIの幾何学的関係のスケッチを図4.1に示す。図の太いベクトルは単語のベクトルを表し、それらが構成する立方体はベクトル空間である。左側の細いベクトルが(P)LSIで求められた軸、右側がSLSIで求められた軸である。両者の違いは(P)LSIで求められた軸には次の制約がある点である。

■軸は直交すること
■元のベクトル空間の内部に存在すること

SLSIは以上のような制約がない事は因子分析を用いた点からでも明らかである。従って潜在的意味の抽出に数学的制約がないので、但し、因子得点の推定に主因子法を用いると、軸の直交という制約が加わることになる。今までの統計的手法の見地からだけでなく、潜在的意味が単語の出現の元になっているという我々の直感からもSLSIが潜在的意味因子を抽出するのにふさわしい事が確認できる。

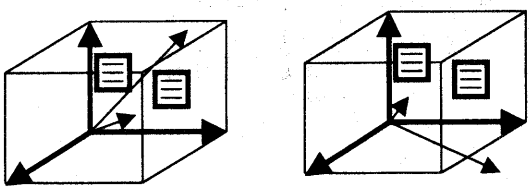

図4.1 (P)LSIとSLSIの幾何学的関係

4 潜在的意味空間における文書の検索

本章では提案手法の有効性を示すために、データとして「情報検索システム評価用テストコレク

ション BMIR-J2」を利用した。提案手法は(P)LSI 同様にキーワード単位の情報に基づいた単語・文書行列を用いるので、本データのうち検索要求のファンクション分類で F1: 基本機能に相当する 14 の検索要求を用いる。データとして利用した新聞記事の数は 5080 あった。これらを茶筅を用いて形態素解析を行い、品詞が名詞あるいは未定義語であるもののみを利用する。その結果、単語の総数は 34902 になり、単語・文書行列は 34902 ×5080 行列となった。

4.1 文書検索

潜在的意味空間の違いによる文書検索の精度比較は次のように行った。まず 14 の検索要求に相当する文書と潜在的意味空間において類似度が高い文書も検索結果に含める事とする。類似度には cosine 関数を用い、その類似度の閾値は 0.60〜1.00 と 0.05 ずつ変化させる。類似度を変化させたときの検索結果の再現率と適合率によって評価を行う。図 1 はその結果を示したものである。図 1 より再現率と適合率が等しくなる提案手法の break-even point が従来の(P)LSI よりも高くなっている。このことは提案手法が抽出した潜在的意味空間の方が精度の高い類似検索が実現できる事が明らかになった。

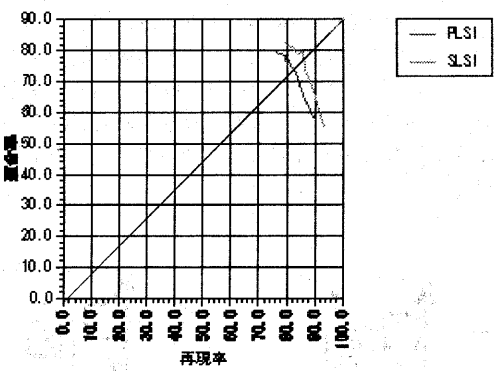

図 5.1 PLSI と SLSI の比較

5 まとめ

本研究は潜在的意味空間を抽出するために SLSI を提案した。SLSI が従来の(P)LSI との大きな相違点は潜在的意味空間と文書・単語の関係をコードモデルと定義し、そのコードモデルを導出するのに因子分析を導入し、目的関数として MDL を導入した点である。実験の結果、因子分析によって抽出された潜在的意味空間の方が文書の検索の精度が良く、目的関数は従来よりも最適な潜在的意味空間の軸の数を決定する事が明らかになった。

参考文献

[1] Chasen.: http://Chasen.aistnara.ac.jp/index.html.ja

[2] Deerwester, S., Dumais, S. T., Furnas, G.W., Landaure, T.K., and Harshman, R.: Indexing by latent semantics analysis, Journal of the American Society for Information Science, 1990.

[3] C.H.Q. Ding.: A Dual Probabilistic Model for Latent Semantic Indexing in Information Retrieval and Filtering In Proceedings of the 22nd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR), 1999.

[4] S.T. Dumais.: Improving the retrieval of information from external sources. Behavior Research Methods, Instruments and Computers, 23(2), 229-236. 1991.

[5] T. Hofmann.: Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR),1999.