データベースシステム 109－2
（1996. 7. 24）

多次元ベクトル空間の視覚的探索機能を有する情報検索

渡辺正裕 石川佳治 吉川正俊 植村俊亮

奈良先端科学技術大学院大学 情報科学研究科

あらゆる分野で莫大な文書が生成，蓄積されているが，本研究では，巨大な文献集合から迅速かつ的確に必要なものを取り出すという要求を満たすために，直観的に情報を提供する手法を提案する．文献-単語行列の特異値分解を利用し，データベースに格納されている文献集合を多次元ベクトル空間として視覚化する．利用者は2次元の入出力インタフェース上で検索目標を指定入力する．その際文献の中で既知であるものとの相対位置関係などを目安に指定することができる．

奈良先端科学技術大学院大学の1994年度の修了者のうち116名の修士論文の英文概要を対象に本論文で提案する手法を適用し，視覚化する実験を行った結果を報告する．

Spatial Indication Mechanism for Exploratory Information Retrieval over Multidimensional Space

Masahiro WATANABE, Yoshiharu ISHIKAWA, Masatoshi YOSHIKAWA and Shunsuke UEMURA

Graduate School of Information Science
Nara Institute of Science and Technology (NAIST)

We propose a new approach to visual information retrieval. In this approach, first of all, a document-term matrix generated from a document set is decomposed using the *singular value decomposition (SVD)* method. Then the document information is visualized over a two dimensional plane using the result of SVD. Users specify retrieval conditions by indicating a position on the plane. With this approach, users can easily specify retrieval conditions which will be complicated when expressed in Boolean expressions of keywords. We report the result of an experiment applying our method to the 116 abstracts of NAIST master theses.

1. はじめに

データベースおよび情報検索の分野において人間-計算機間の入出力インタフェースを研究することの重要性が指摘されている [Hab94]. 特に, 計算機の能力が向上したことにより, 3次元グラフィクスや仮想現実感を利用して計算機の出力を人間にとって理解しやすいように視覚化し, 直接操作出来るようにするといった研究が盛んに行なわれている [Kei96, Dub95, BM94].

筆者らは, 問合せおよびデータの視覚化について研究を進めている [渡辺 95, 渡辺 96]. 本研究では文献集合を対象とし, 各文献の内容を概念ベクトルとして表現するモデルを前提とする. その上で検索条件の入力や検索結果の出力を視覚的に行なう情報検索の新しいアプローチを提案する.

本論文では, 文献-単語行列の特異値分解を行い, その結果得られる文献空間の特徴ベクトルから検索インタフェース平面を構成する手法を導入する. さらに提案する手法を実際に奈良先端科学技術大学院大学の情報科学研究科の博士前期課程修了者のうちの 116 名の修士論文について適用する実験を行い, 本手法が有効かどうかを検証する.

検索の入出力を2次元インタフェース上で行う本アプローチでは, キーワードや, 利用者にとって既知のオブジェクトを2次元平面上に配置することにより, 検索目標との相互位置関係によって視覚的に検索条件を指定する. 検索結果として検索目標の近傍のオブジェクトが視覚化して提供される. この過程を繰り返すことにより情報検索作業を進める. このような情報検索アプローチでは, キーワードのブール式などでは表現が困難な複雑な条件を視覚的かつ直観的に指定することができ, また検索条件の微妙な調整も行ないやすい. 視覚的な検索条件の指定によってシステムと利用者の対話はより円滑なものになる.

2. 多次元ベクトル情報の処理手法

本研究は, 対話的な情報検索手法の提案を目的としている. システムの応答が遅いと, 利用者が満足する検索結果を得るまで対話を続けるのに根気がいるため, 途中で検索をあきらめてしまう傾向がある. そのため, 効果的な検索を実現するためには1回の応答に時間があまりかからないことが必要である.

多次元ベクトル情報を視覚化するために採用する指針としては, データが散らばることと, 計算量ができるだけ小さいことがあげられる. 前者の要求は利用者がシステムから提供されたデータを一見して理解, 判断, 迅速に次の検索行動に移れ, 検索目標の指定がしやすいことが重要であると考えられるためである. 後者の要求は, 巨大な文献集合が利用可能になってきており, 計算量を少しでも小さくする必要があるためである.

本節では多次元ベクトル情報をもとに, 文献空間の構造を少ないパラメータで近似的に表現し, その構造を視覚化するのに利用可能な手法をいくつか述べる.

2.1 *K-L* 展開

ベクトル情報を行列で表現したときに, そこからどのようにして視覚化のための情報を取り出すかを扱う手法として, *K-L* 展開 (*K-L* 変換) [森 90, FL95] が知られている. *K-L* 展開は, パターン認識の分野では古典的な手法であり, データが与えられると, これらを正射影した時に分散が最大になるように直交座標系を作りだす. つまり, データがもっとも散らばって見える方向に射影する. *K-L* 展開の通常の方法ではベクトル数 N に対して $O(N^2)$ の時間計算量が必要であり, 大規模なデータベースに適していないなどの問題点が存在する. これを近似的に線形時間で解決するために提案されたのが Faloutsos らによる *FastMap* アルゴリズム [FL95] である.

2.2 特異値分解

特異値分解は主成分分析, パターン認識, 逆問題分析など多くの分野で用いられる手法である [岡本 92]. この特異値分解を文献集合に適用する方法について述べる. 文献集合から, 文献-単語行列を生成する (図 1). この行列の各成分はどの文献にどの単語が何回出現しているかを表す整数であったり, 出現しているかどうかの2値であったりする. この行列の特異値分解 (Singular Value Decomposition ; SVD) を計算する (図 2).

文献-単語行列 X_0 は t 行 (文献集合に出現する各単語に対応) と, d 列 (集合中の各文献に対応) から構成される. SVD $X_0 = T_0 S_0 D_0^T$ は,

- $t \times m$ 行列 T_0 : 左特異ベクトル行列と呼ばれる列直交行列 (図 3左)
- $m \times m$ 行列 S_0 : 対角要素が降順に並べられた正の特異値 (singular value) であるような対角行列 (図 3右)

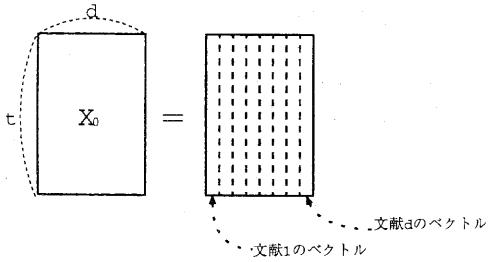

図 1 文献-単語行列 X_0

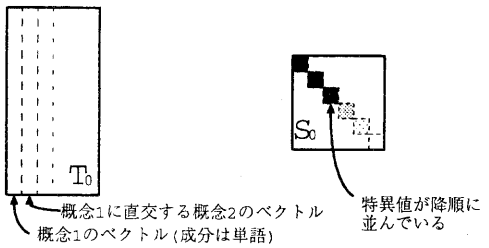

図 3 X_0の特異値分解の結果の行列

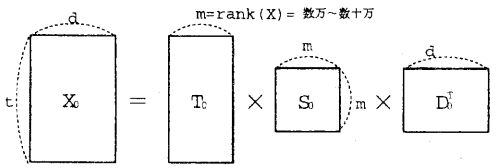

図 2 X_0 の特異値分解

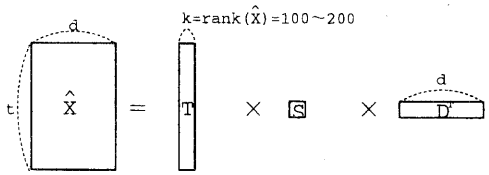

図 4 $\hat{X}$ の特異値分解

- $m \times d$ 行列 D_0：右特異ベクトル行列と呼ばれる列直交行列

に帰着する．T_0，S_0，D_0によってX_0を正確に再現できる．ベクトル空間内の特徴を扱う場合には各行列の小さい特異値に対応する部分は，不要であるばかりかノイズにもなるので，これを無視して近似した行列$\hat{X}$を扱うことが多い (図 4)．そのとき，行列の階数が m から kに削減される．これは，行列に非常に小さい特異値が存在して丸め誤差の影響が大となるような場合に場合に，安定した性能を発揮するようにするための一般的な手法である [PTVF88]．

文献と文献の間の相関を調べる際に，自己相関行列 $X_0^T X_0$を求めることが一般に行われるが，これはX_0の次元数を削減した行列$\hat{X}$の自己相関行列

$$\hat{X}^T \hat{X} = DS^2 D^T$$

を求めることで近似できる．同様に，与えられた文献集合中での単語と単語の相関を

$$\hat{X}\hat{X}^T = TS^2 T^T$$

で求めることができる．$\hat{X}$の次元数が小さいことより，X_0について計算する場合に比べこのような計算を効率的に処理することができる．問合せに対する文献のランキングや文献集合のクラスタリングなどにおいて，余弦類似尺度の計算や文献ベクトル間の距離の計算が一般に求められるが，次元数を削減した行列$\hat{X}$を利用することにより，少ない計算量で近似的な値を求めることが可能である．

3. 文献検索インタフェースへの応用

本節では文献-単語行列を特異値分解を文献検索に応用する方法について述べる．文献検索に用いる問合せは，数個の単語から構成され，m 次元ベクトル qとして表現できる．図4のように削減した行列のもとで，各文献のベクトルと比較するために，

$$\hat{q} = q^T T S^{-1}$$

としてk次元のベクトルで表現することが必要である．

3.1 単語を軸とした 2 次元インタフェース

本節では [渡辺 96] で提案している 2 次元インタフェースへの応用の方法を示す．このインタフェースでは，図 5 に示すように，利用者が選択した二つの単語との相関関係に基づいて文献が 2 次元平面上に表示される．利用者はインタフェース上に表示された文献を直接指示して検索したり，2 次元平面上である点を指定してその近傍の文献を検索することができる．

その実現のためには，前節で述べた特異値分解の

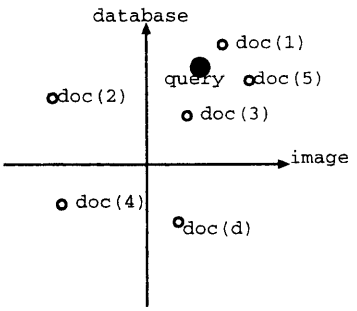

図 5 単語を軸とした2次元インタフェース

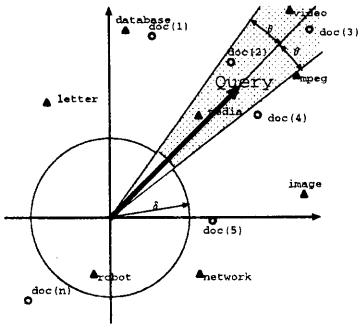

図 6 $\hat{X}$の特徴ベクトルを軸とした2次元インタフェース

結果得られる行列$\hat{X} = TSD^T$を利用する．図4に示すように，$\hat{X}$は元の文献-単語行列 X_0と同様に文献と単語の間の関連を示している．しかし，$\hat{X}$は特異値分解の結果からノイズに相当する特異値を削減し再び合成したものであるため，X_0に比べ文献と単語の間の関連をより的確に表現しているのではないかと考えられる．

このインタフェースでは，利用者が指定した単語t_a，t_bのそれぞれについて$\hat{X}$から対応する行ベクトルを抜きだし，その情報を2次元平面に表示する．利用者が2次元平面上の一点を示し，その近傍を取り出すよう指示したとき，システムは余弦類似尺度を計算することにより，近傍の文献を求め利用者に提供する．

3.2 特徴ベクトル空間に対する2次元インタフェース

本節では$\hat{X}$の各特徴ベクトルを軸とした2次元平面インタフェースへの応用の方法を示す．文献-単語行列$\hat{X}$の特異値分解で得られるTとSにおいて，文献と単語に対応するk次元ベクトルの各成分は互いに直交する特徴成分を表現している．これらのうちの2成分を取り出してそれらの値を横軸と縦軸に対応させると文献と単語のベクトルを2次元に配置することができる (図 6)．利用者はこの2次元インタフェース上で問合せベクトルを指定する．このとき指定しやすいような2次元配置を得るためには，配置の目安となるいくつかの単語や利用者にとって既知の文献のベクトルがインタフェース上に散らばっていることが必要である．したがってこれらの目安となる単語や文献のベクトルの分散が大きくなるように2次元の軸を選択する．

この座標軸上で，問合せベクトルとのなす角が小さい (θ以下の) 文献ベクトルを検索結果の候補とする．しかし，ベクトルの絶対値が小さい (δ以下の) 文献はこの座標軸では特徴がよく表現されないので検索結果には含めない．

4. 実験と評価

奈良先端大の情報科学研究科の 1994 年度の修了生のうち 116 人の修士論文の英文概要を対象に，本論文で提案している手法を適用した実験について報告する．

4.1 手順

1. 116 人分の修士論文の英文概要を文献集合全体とする．
2. 各文献に語幹の切り出し処理 (stemming) を施し，不要語 (stop words) を除去する [FBY92].
3. 文献集合内で用いられている異なり単語 (different term) の数を数え，文献-単語行列を作る．今回の実験では行列の成分に各単語が存在するかどうかの2値を採用した．
4. 行列の特異値分解を行う．
5. 2次元平面座標を構成する二つの特徴ベクトルの組合せを選択する．このとき利用者が二つの単語，t_1とt_2を目安にして検索目標ベクトルを指定しようとする場合考えると，これらの単語がつぎのように配置されていることが望ましい．
   - t_1とt_2のベクトルが原点からある程度距離を持つ
   - t_1とt_2のベクトルが原点を頂点としてなす角が大きすぎず，また小さすぎない

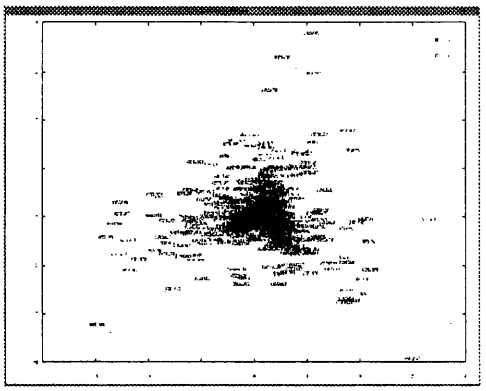

図 7 X_0の成分を軸として2次元配置された単語と文献の例

これらの要求を満たすようにするためには，一方の単語は横軸の成分だけが，他方の単語は縦軸の成分だけが大きくなっているように座標軸の組合せを選択すれば良い．こうすることで利用者の指定の意図をよく反映した2次元平面で表示することができると考えられる．今回は次のような基準で軸となる二つの特徴ベクトルの組合せを選択した．

- 対応する特異値が大きいものの上位 50 番目までである
- その軸で配置したときにt_1とt_2のベクトルの大きさがともに 0.1 より大きい
- その軸で配置したときに t_1と t_2のなす角の余弦が最小になる (つまり，2次元平面に射影したときに t_1と t_2が離れて表示される)

4.2 結果

図 7に修士論文の概要を 2 番目と 3 番目の成分で構成される平面に 2 次元配置した例を示す．この座標軸の組合せでは,「object」と「database」の両ベクトルに挟まれた領域にオブジェクト指向データベースに関する文献が目立って見受けられた．

4.3 考察

今回の実験では 4.1節の 5で述べた指針に従って「database」「image」について軸を選択した．「database」と「image」の中間の方向を持ったベクトルに対応する文献が動画像データベースに関する文献であることを期待したが，実際には動画像データベースに関する文献は予想と異なる方向のベクトルに対応していた．

期待通りの結果が得られなかった原因としては次のようなことが考えられる．

- 各文献が修士論文の英文概要であり，非常に短い
- 文献集合の要素が少なすぎる
- 文献の特徴を表現するのには 2 次元では少ない

これらの要素について実験を行い，検証することが今後の課題である．

また，座標軸の組合せをさまざまに変化させて「database」と「image」を配置してみると，座標軸のとりかたによっては利用者の期待どおりの配置になる場合があり，そのような場合は利用者の視点で検索目標を指定できる可能性があると考えられた．そこで，

1. 利用者は 4.1節の 5で述べた指針において，t_1とt_2の内積が小さくなる順に座標軸に採用する特徴ベクトルの組合せを変えて文献を配置してみる．
2. 利用者にとって既知の文献が予想通りに配置されるような特徴ベクトルの組合せになるまで特徴ベクトルの組合せを試みる．

といった方法で検索目標ベクトルに思い通りの検索意図を反映させることができるようなシステムの可能性について検討をする必要がある．

5. 関連研究

文献集合の文献-単語行列を用いて文献検索に利用する研究がいくつか行われている．これらのうちのいくつかを紹介する．

Latent Semantic Indexing (LSI) [DDF+90, BDL95] は特異値分解を用いたベクトル空間情報検索手法であって，[FO95] によれば，Salton の SMART システム [SM83] で用いられた伝統的なベクトル空間技術を上回る性能を発揮する．

これらの文献ベクトルと問合せベクトルの内積を用いて余弦類似尺度 (cosine measure) を計算することができる．これをもとに文書は問合せとの類似度でランク付けされ，その順番に利用者に返される．

LSIの特長は，文献-単語行列の特異値分解の結果に対して，小さな特異値に対応する次元の削除を行い，計算コストの削減とノイズの除去を行うことである (図4)．LSIの手法は近年計算機の性能の向上によって大規模な文献-単語行列の特異値分解の計算が可能になったことから注目されている．

清木らの「意味の数学モデル」[KKH94, 北川94]は文献-単語行列 X_0 について相関行列 $X_0^T X_0$ の固有値分解を行う点は本研究およびLSIに関連が深い．

本研究では特異値分解を情報の視覚化に用いた点が特徴である．

6. 今後の課題

今回行った実験では，利用者が目安にしたいと考える単語を恣意的に選択することにした．しかし，目安にする単語に何らかの制限を与えることによって，これらの単語に基づいて文献集合を2次元平面に分散して配置できれば理想的である．[SM83, Dub95]は文献のクラスタを識別する単語の性能 (group discriminatory power) について議論している．いわゆる"good discriminator" を本研究の座標軸の選択基準に用いて，文献を分散して表示する性能を改良できないか検討することは今後の課題のひとつである．

謝辞

本研究を進めるにあたって日頃から有意義な御指導・御討論をいただく植村研究室の皆様に感謝いたします．

参考文献

[BDL95] M.W. Berry, S.T. Dumais, and T.A. Letsche: "Computational Methods for Intelligent Information Access", in *Proceedings of Supercomputing*, 1995.

[BM94] Steve Benford and John Mariani: "Virtual Environments for Data Sharing and Visualizaion – Populated Information Terrains", in *Interfaces to Databese Systems*, pp. 168–182, 1994.

[DDF+90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman: "Indexing by Latent Semantic Analysis", in *Journal of the American Society for Information Science*, Vol. 41-6, pp. 391–407, 1990.

[Dub95] David Dubin: "Document Analysis for Visualizaion", in *ACM SIGIR '95*, pp. 199–204, 1995.

[FBY92] William B. Frakes and Ricardo Baeza-Yates, editors: *Information Retrieval – Data Structures & Algorithms* –, Prentice-Hall, 1992.

[FL95] Christos Faloutsos and King-Ip(David) Lin: "FastMap: A Fast Alghorithm for Indexing, Data-Mining and Visualisation of Traditional and Multimedia Datasets", in *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 163–174, 1995.

[FO95] Christos Faloutsos and Douglas W. Oard: "A Survey of Information Retrieval and Filtering Methods", Technical Report CS-TR-3514, University of Maryland, Aug. 1995.

[Hab94] Eben M. Haber: "The Ambleside Surbey: Important Topics in DB/HCI Research", in *Interfaces to Database Systems, Lancaster*, pp. 361–364, 1994.

[Kei96] Daniel A. Keim: "Databases and Visualization", Jun. 1996, Tutorial Notes in SIGMOD96.

[KKH94] Yasushi Kiyoki, Takashi Kitagawa, and Takanari Hayama: "A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning", in SIGMOD RECORD, Vol. 23–4, pp. 34–41, Dec. 1994.

[PTVF88] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery: *Numerical Recipes in C*, Cambridge University Press, 1988.

[SM83] G. Salton and M. McGill: *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[岡本92] 岡本 良夫：逆問題とその解き方，オーム社，1992年．

[森90] 森 俊二, 坂倉 栂子：画像認識の基礎[II]，オーム社，1990年．

[渡辺95] 渡辺 正裕, 吉川 正俊, 植村 俊亮："対話的質問作成と提案機構に基づくデータベース利用者インタフェース"，情報処理学会 第51回全国大会，2D-03，1995年9月．

[渡辺96] 渡辺 正裕, 吉川 正俊, 石川 佳治, 植村 俊亮："視覚的対話機能を有する情報検索インタフェース"，情報処理学会データベースシステム研究会研究報告，第96-DBS-106巻，pp. 9–16，1996年1月．

[北川94] 北川 高嗣, 清木 康, 人見 洋一："文脈理解機能をもつ意味の数学モデルのデータベース処理への適用"，Technical Report DE94-4，信学技報，1994年5月．