2005－DBS－137（Ｉ）（4）
2005／7／13

ドキュメント上の単語分布に基づく検索空間の生成方式

本間　秀典 †　　中西　崇文 †　　北川　高嗣 ††

† 筑波大学大学院 システム情報工学研究科　〒305–8573 茨城県つくば市天王台 1–1–1 情報数理研究室
†† 筑波大学大学院 システム情報工学研究科　〒305–8577 茨城県つくば市天王台 1–1–1
E-mail: †{homma,takafumi}@mma.cs.tsukuba.ac.jp, ††takashi@cs.tsukuba.ac.jp

概要　ある特定の分野に関して意味の数学モデルによる連想検索を実現するためには，その分野を対象としたメタデータ空間と呼ばれる検索空間を生成する必要がある．これまで，検索空間は，辞書や用語辞典，専門的な知識などを用いて生成していたが，これらの方法が利用できない場合，検索空間の生成が困難であった．本論文では，ドキュメント中における単語間の関連性を反映した連想検索のための検索空間の生成方式を提案する．提案方式は，ドキュメント上の単語の分布の傾向に注目し，単語間の関連の度合いを計算することによって検索空間を生成する．提案方式により，意味的連想検索を実現するための検索空間を，辞書や用語辞典などを用いることなく自動的に生成することが可能となる．

A Construction Method of a Retrieval Space Based on Word Distributions in Documents

Hidenori HOMMA†, Takafumi NAKANISHI†, and Takashi KITAGAWA††

†Graduate School of Systems and Information Engineering, University of Tsukuba
††Graduate School of Systems and Information Engineering, University of Tsukuba
E-mail: †{homma,takafumi}@mma.cs.tsukuba.ac.jp, ††takashi@cs.tsukuba.ac.jp

Abstract　In order to realize associative search for a specific field by the mathematical model of meaning, it's necessary to construct a retrieval space. A construction method of retrieval space has been proposed that creates a data matrix by using a dictionary or a term dictionary. However, it's difficult to construct a retrieval space without them. This paper presents a new construction method of a retrieval space based on the relation between each word in documents. This method constructs a retrieval space by the relevance between each word calculated from word distributions in documents. By this method, a construction of a retrieval space for an associative search is performed automatically without any dictionary.

1　はじめに

コンピュータネットワーク上に様々な特定分野に関する多種多様な情報群が蓄積されつつある．これらの情報の種類もドキュメントデータをはじめ画像，音楽，音声，動画など様々であり，これらを対象とした高度かつ効率的な情報検索方式，および知識の獲得方式の必要性が高まってきている．近年，このような情報群を対象とした高度な検索方式としてベクトル空間検索方式が提案され，その効果が確認されている．ベクトル空間検索方式とは，検索対象を表現するベクトルを検索空間と呼ばれるベクトル空間に配置し，その空間内に検索質問に対応するベクトルとの計量によって情報検索を行う方式である．

これまで，検索空間を用いた言葉と言葉の関係の計量による検索機構として，意味の数学モデルが提案されている [1, 2, 3]．これは，単語群を文脈として解釈する機構により，言葉と言葉，あるいは，言葉と検索対象のメディアデータ，ドキュメント間を文脈に応じて動的に計算することを可能とする．意味の数学モデルでは，検索対象をベクトル化して検索空間に写像し，それらのベクトルを検索空間の部分空間に射影して計量することにより，文脈に応じ

た連想検索を実現している．ここで，検索空間は，基本データと呼ばれる特徴付きベクトルの集合であるデータ行列から生成される．このデータ行列の生成方式として，辞書や用語辞典において説明される言葉（見出し語）の語義文中で見出し語の説明に使われている語（特徴語）を用いてその見出し語の特徴付けを行う方式が提案されている [2, 4]．しかしながら，これらの方式では辞書や用語辞典から適切に特徴付けを行うことができない場合に適用が困難であることが問題であった．

そこで本研究では，辞書や用語辞典を用いる代わりに，検索対象を包含する任意のドキュメントを用いて検索空間を生成する方式を提案する．本方式によって単語間の関連性を計量する検索空間を生成できれば，単語間の関連に基づく連想検索，すなわち，単語間関連連想検索の実現が可能になると考えられる．本方式には，対象とする分野に関するドキュメント中に出現する各単語の距離と頻度を用いて検索空間を構成するため辞書や用語辞典などを必要とすることがなく，さらに検索空間を生成するためのデータ行列の作成を自動化できるという特徴がある．さらに，本方式で生成されたメタデータ空間を意味の数学モデルへの適用実験についても示し，単語間関連連想検索の実現を目指す．

2 ドキュメント上の単語分布に基づく検索空間の生成方式

本節では，特定分野に関するドキュメントから検索に利用する言葉を自動抽出し，さらにそれらの間の相関を用いたデータ行列を自動生成するための方式について述べる．なお，ここでは対象となる分野に関するドキュメントが存在し，かつ何らかの方法で検索対象となるメディアからメタデータを抽出できることを前提としている．

2.1 ドキュメント中に出現する単語間の関連性

言語学者 Zellig Harris は，単語や形態素の意味の違いはそれらの分布の違いと相関があり，さらに，そのような分布的な性質は狭い範囲において見出すことができる，と述べている [5]．このため，単語が出現する場合の距離が小さく，かつその頻度が大きい単語ほど意味的に大きな関連性を持っていると考えられる．そこで，ある概念を説明するために書かれたドキュメント内に出現するある語 w_1 とその近辺に出現する語 w_2 の関係について，次のような性質があると考えられる．

- ある単語 w_1 の近辺に出現する幾つかの単語がある場合，その出現位置が w_1 から近いものほど関連性が強い．
- ある単語 w_1 が同一ドキュメント内に複数回出現し，かつ w_1 から等しい距離に出現する単語が複数ある場合，出現する頻度の高いものほど関連性が強い．

単語間の関連に関するこれらの性質は，ドキュメント中に出現する各単語間の関連を求める上で重要であると考えられる．この性質を用いることによって対象となる分野に関するドキュメントから抽出した単語間の距離とその出現する頻度をもとにメタデータ空間を生成し，単語間の関連性に基づく連想検索である単語間関連連想検索を実現できると考えられる．

2.2 単語間の関連性の定量化

はじめに，単語間の距離と，距離に基づく重みを設定する．まず，隣接して出現する 2 語間の距離を 1 とする．このとき，2 語が間に r 語を挟んで出現する場合の単語間の距離を d とすると，$d = r + 1$ となる．この d を用いて，2 語が隣接する場合を 1 とし，距離が大きくなるにつれて関連度が大きく下がるような評価関数 $W(d)$ を設定する．ここでは，予備実験により $W(d)$ を式 (1) のように設定している．

$$W(d) = \mathrm{e}^{1-d} \tag{1}$$

次に，式 (2) により単語 w_j が単語 w_i から距離 d の位置に出現する頻度 $P_{ij}(d)$ を求める．

$$P_{ij}(d) = \frac{(w_j\text{が } w_i\text{から距離 } d \text{ に出現した回数})}{(w_i\text{がドキュメント内で出現した回数})} \tag{2}$$

ここで，対象となるドキュメントが N 語の語列からなるとすると，式 (1)，(2) により，単語 w_i と w_j

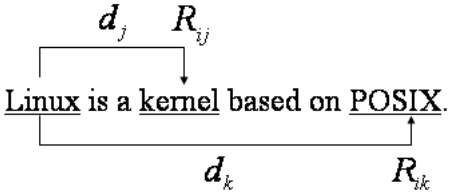

図 1　距離重み頻度関数の適用例

の関連を表す関数 R_{ij} は式 (3) のように計算できる．

$$R_{ij} = \sum_{d=1}^{N-1} P_{ij}(d) \times W(d) \tag{3}$$

R_{ij} は w_i と w_j の出現頻度と出現時の距離に依存する関数であるから，「距離重み頻度関数」と呼ぶことにする．ただし，

$$w_i w_i w_i \ldots w_i$$

のように，単一の単語 w_i のみからなる N 語の語列を考えると，$d = 1, 2, \ldots, N$ に対して $P_{ij}(d) = 1$ は明らかであり，しかも 3 つ以上同じ語が連続しても意味があるとは考えにくい．このとき，

$$\sum_{d=1}^{2} P_{ii}(d) \times W(d) \leq \sum_{d=1}^{2} W(d) \tag{4}$$

は明らかであるから，w_i と w_i の距離重み頻度関数値は式 (5) のように与えられるものとする．

$$R_{ii} = \sum_{d=1}^{2} W(d) \tag{5}$$

ドキュメント中に出現する言葉と言葉への距離重み頻度関数 R_{ij} の適用例を図 1 に示す．

2.3　ドキュメント中に出現する単語間の関連性を反映した検索空間の生成

ここでは，対象となるドキュメント中に出現する単語間の関連性を用いた検索空間の自動生成方式を示す．その具体的な手順は以下のようである．

1. ドキュメントの解析
   対象となるドキュメントから検索に利用する単語のみを抽出する．この作業は形態素解析器や単語抽出ツールなどを用いることによって自動化することができる．本論文では，日本語形態素解析器『茶筌』[6]，及び東京大学中川研究室・横浜国立大学森研究室で開発された『専門用語自動抽出システム』[7] を利用して単語の抽出を自動化している．
2. 単語間の関連性を反映したデータ行列の作成
   手順 1 で得られた N 語の語列から重複して出現する単語を除いた語数を n とする．ここで，2.2 節に示した方式によって各単語間の R_{ij} の値を計算することにより，(6) 式のように単語 w_i を特徴付きベクトルとして表すことができる．

$$\boldsymbol{w}_i = (R_{i1}, R_{i2}, \cdots, R_{in}) \tag{6}$$

   この $\boldsymbol{w}_i$ を用いて $(\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n)^{\mathrm{T}}$ とすることによって，図 2 のような n 次正方行列 M を作成する．

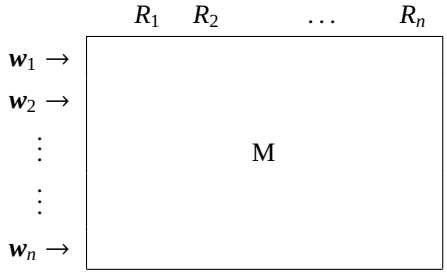

図 2　データ行列 M の表現

3. 相関行列 $M^{\mathrm{T}}M$ から検索空間生成
   手順 2 で作成されたデータ行列 M の相関行列 $M^{\mathrm{T}}M$ を計算すると n 行 n 列の行列となる．この相関行列 $M^{\mathrm{T}}M$ を固有値分解し，非ゼロ固有値に対応する固有ベクトルによって検索空間を生成する．これにより，語と語の関係を計量する単語間関連連想検索のための検索空間の構成が可能となる．

3　意味の数学モデルへの適用

本節では，2 節で生成された検索空間を意味の数学モデルに適用することにより，単語間関連連想検索の実現方法を示す．意味の数学モデルの詳細は，文献 [1, 2, 3] に示している．

1. 検索対象データのメタデータを検索空間へ写像

   メタデータ空間へ検索対象データのメタデータをベクトル化し写像する．これにより，検索対象データが同じ検索空間上に配置されることになり，検索対象データ間の関係を空間上での語と語の関係として計算することが可能となる．

   検索対象データ D には，メタデータとして t 個の語 $o_1, o_2, \cdots, o_t$ が以下のように付与されていることを前提としている．

$$D = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_t\}. \tag{7}$$

   ここで，各印象語 $\boldsymbol{o}_i$ は，データ行列の特徴語と同一の特徴を用いて表現される特徴付ベクトルである．

$$\boldsymbol{o}_i = (o_{i1}, o_{i2}, \cdots, o_{in}) \tag{8}$$

   各検索対象データは，メタデータとして付与されている t 個の語が以下のように合成され，検索対象データベクトル $\boldsymbol{d}$ を形成する．

$$\begin{aligned}\boldsymbol{d} &= \bigoplus_{i=1}^{t} \boldsymbol{o}_i \\ &:= (\mathrm{sign}(o_{\ell_1 1}) \max_{1\le i\le t} |o_{i1}|, \\ &\qquad \mathrm{sign}(o_{\ell_2 2}) \max_{1\le i\le t} |o_{i2}|, \\ &\qquad \cdots, \mathrm{sign}(o_{\ell_n n}) \max_{1\le i\le t} |o_{in}|).\end{aligned} \tag{9}$$

   この和演算子 $\bigoplus_{i=1}^{t}$ は，t 個のベクトルから各基底に対して絶対値最大の成分を選ぶ演算子である．ここで sign(a) は，“a” の符号（正，負）を表す．また，$l_k (k = 1, \ldots, t)$ は，特徴が最大となる印象語を示す指標であり，次のように定義する．

$$\max_{1\le i\le t} |o_{ik}| = |o_{\ell_k k}|. \tag{10}$$

   これにより検索対象データのメタデータがデータ行列の特徴語と同一の特徴を用いて表現される．検索対象データベクトル $\boldsymbol{d}$ を検索空間へ写像する．この写像は，検索対象データベクトル $\boldsymbol{d}$ をメタデータ空間内でフーリエ展開し，フーリエ係数を求める．

2. 検索空間の部分空間の選択と相関の定量化

   検索者が与える単語の集合をコンテクストと呼ぶ．コンテクストを用いて検索空間に各単語に対応するベクトルを写像する．これらのベクトルは検索空間において合成され，意味重心を表すベクトルが生成される．意味重心から各軸への射影値を相関とし，閾値を超えた相関値を持つ軸からなる部分空間が選択される．選択された検索空間の部分空間において，検索対象データベクトルのノルムを検索語列との相関として計量する．これにより検索者が与えた検索語と各ドキュメントデータとの相関の強さを定量化する．この部分空間における検索結果は，各検索対象データを相関の強さについてソートしたリストとして与えられる．

4　実験

提案方式により生成される検索空間の性質を確認するため，提案方式に基づくデータ行列作成システムを構築し，実験を行った．

4.1　実験環境

本実験において，検索空間を生成するためのドキュメント群として，Web サイト “@IT”[8] から Linux 基礎事項に関する記事 22 ページを利用した．また，提案方式における特徴語の抽出方式として，専門養護児童抽出システム [7] を採用した．

4.2　実験方法

上記のドキュメント群に提案方式を適用して検索空間を生成し，さらに生成された検索空間を意味の数学モデルに適用して検索実験を行った．本実験では，抽出された特徴語を検索語，および検索結果として用いた．これにより，提案方式により生成される検索空間上における言葉と言葉の関連を確認

表 1　実験結果 (コンテクスト： パーティション)

パーティション	0.564136
HDD	0.481881
Mbytes	0.433038
ファイルシステム	0.383270
スワップ領域	0.347455
Gbytes	0.328363
ディレクトリ	0.321301
BIOS	0.319955
残り	0.308639
PC	0.306166

表 2　実験結果 (コンテクスト： サーバ)

サーバ	0.575058
アタック	0.523818
SPAM メール	0.521262
インターネット	0.503491
ホスト	0.500154
LAN	0.481603
勉強	0.461589
踏み台	0.441388
注意点	0.388292
マシン	0.384502

表 3　実験結果 (コンテクスト： vi)

文字	0.593283
カーソル	0.573758
一致	0.498775
vi	0.483039
キー	0.482889
単語	0.450396
入力	0.411259
ESC	0.410312
挿入	0.408746
削除	0.393274

した.

4.3　実験結果

検索語を「パーティション」,「サーバ」, および「vi」とした場合の検索結果の上位 10 件とその相関量をそれぞれ表 1, 2, および 3 に示す.

4.4　考察

表 1 では,「パーティション」というコンテクストから,「HDD」や「Mbytes」など, ハードディスクの領域や容量などに関する語が上位に検索されていることが分かる. また, 表 2 では,「アタック」「SPAM メール」「踏み台」といったセキュリティ関連の言葉や, ネットワークに関連する言葉である「インターネット」「ホスト」「LAN」のように,「サーバ」というコンテクストから連想される言葉が上位にランクされていることが確認できる. 表 3 のコンテクスト「vi」はテキストエディタの名称であるが, 検索結果を見てみると,「カーソル」「入力」「ESC」「挿入」「削除」や「文字」「一致」など, 上位にランクされた言葉はすべてテキストエディタの機能や操作方法に関すると考えられる言葉である.

これらの実験結果から, 提案方式によって, ドキュメントの内容から言葉と言葉の関連を計量し, 検索空間に反映することができたと考えられる.

5　おわりに

本論文では, ドキュメント中に出現する単語間の関連性を反映した連想検索のための検索空間の生成方式を提案した. 本方式は, 対象となる分野に関する適切なドキュメントを選択し, そのドキュメント中における各単語間の距離や頻度などの分布からそれらの間の関連度を評価することによって, 辞書や用語辞典を必要とせず, かつ自動的に検索空間を生成することができる. また, 本方式により生成される検索空間を意味の数学モデルに適用することにより, 単語間の関連に基づく連想検索である, 単語間関連連想検索を実現することが可能になると考えられる.

今後の課題として，本方式の定量的な評価や，様々な特定分野への適用実験などが挙げられる．また，複数のドキュメント群から生成される検索空間の連携や統合を行うことによりさらに高度な検索システムを構築できると考えられるため，こうした複数の検索空間の連携，および統合方式についても考察すべきである．

参考文献

[1] Kitagawa, T. and Kiyoki, Y.: The mathematical model of meaning and its application to multidatabase systems, Proceedings of 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp. 130-135(1993).

[2] Kiyoki, Y., Kitagawa, T. and Hayama, T.: “A Metadatabase System for Semantic Image Search by a Mathematical Model of Meaning,” Multimedia Data Management – using metadata to integrate and apply digital media –, McGrawHill, A. Sheth and W. Klas(editors), Chapter 7 (1998).

[3] 清木康, 金子昌史, 北川高嗣: “意味の数学モデルによる画像データベース探索方式とその学習機構, ” 電子情報通信学会論文誌,D-II,Vol.J79-D-II,No. 4,pp. 509-519 (1996).

[4] 宮川祥子，清木康: “特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式 ,” 情報処理学会論文誌：データベース, Vol.40, No.SIG5(TOD2), pp.15-27,(1999).

[5] Zellig S Harris / edited by Henry Hiz. Reidel：“Papers on syntax” (1981)

[6] http://chasen.naist.jp/

[7] http://gensen.dl.itc.u-tokyo.ac.jp/index.html

[8] http://www.atmarkit.co.jp