情 報 学 基 礎 53－1
（1999．3．1）

単語の連想関係に基づく情報検索システムＩｎｆｏＭＡＰ

高山 泰博*, Raymond Flournoy**, Stefan Kaufmann **, Stanley Peters**

*三菱電機（株）情報技術総合研究所
**スタンフォード大学 言語・情報研究センター

E-mail: takayama@isl.melco.co.jp
{flournoy, kaufmann, peters}@csli.stanford.edu

情報検索の対象となるテキスト文書は単語の集まりから構成されており、単語の意味表現は情報検索システムの基礎として重要である。本研究では、コーパスにおける単語の共起関係から導出した多次元の単語ベクトルで単語の意味を表現する。単語ベクトルをアプリケーションで利用する場合には、ベクトルの次元の大きさが問題となる。我々は単語ベクトルの次元を縮退させるために特異値分解(SVD)を用いる。また、単語ベクトルを文脈ベクトルに拡張して InfoMap と呼ぶ情報検索システムを構築した。本稿では、SVD を用いた単語ベクトルの構築方法、SVD と主成分分析との関係、InfoMap 情報検索システムの構成と分野依存の英語コーパス OHSUMED を用いた予備的な実験結果について報告する。

An Information Retrieval System based on Word Associations - InfoMap

Yasuhiro Takayama*, Raymond Flournoy**, Stefan Kaufmann**, Stanley Peters**

*Information Technology R & D Center, Mitsubishi Electric Corporation
**Center for the Study of Language and Information, Stanford University

E-mail: takayama@isl.melco.co.jp
{flournoy, kaufmann, peters}@csli.stanford.edu

The text documents targeted for Information Retrieval (IR) consist of the collection of words, so representation of word senses is an essential basis for IR systems. We have developed representation of word senses as word vectors derived from text corpora. We have to concern the dimensionality in order to use word vectors for applications, so we use Singular Value Decomposition (SVD) as a dimensionality reduction tool. Furthermore we extend word vectors to context vectors and construct IR system called InfoMap. This paper describes how to create word vectors, the relation between SVD and Principal Component Analysis, the configuration of InfoMap system, and a preliminary experimental result using a domain-specific English text corpus OHSUMED.

1．はじめに

情報検索の対象となるテキスト文書は、単語の集まりから構成されており、単語の意味表現の方法は、概念ベースの情報検索システムの基礎として重要である。そこで、単語の意味を多次元の単語ベクトルとして表現する研究が行なわれている。この単語ベクトルの基底とする情報には、辞書に基づくものとコーパスの共起関係に基づくものがある[丹羽 93]。このいずれの場合でも、これらの単語ベクトルを情報検索等のアプリケーションに応用する場合には、ベクトルの次元の大きさが問題となる。

辞書に基づいて単語間の意味距離を計算する手法に小嶋らの研究[小嶋 97]がある。小嶋らは、活性伝播と呼ばれる手法で英語辞書LDOCE(Longman Dictionary of Contemporary English)の定義語間の重みによるベクトル空間を計算し、主成分分析によってベクトルの次元を縮退させて単語の意味空間を構成している。しかし、分野依存の文書データを扱う場合には、LDOCEのような辞書の存在を期待できない。

そこで、我々はコーパスにおける単語の共起ベクトルを基に単語ベクトルを作成する手法を検討した。共起ベクトルの次元を縮退させるために、特異値分解(SVD: Singular Value Decomposition)を用いて Word Space と呼ぶ単語の意味空間を構成した。本稿では、Word Space の構築方法について述べ、SVD の計算が多変量解析における主成分分析とほぼ等価な手法であることを説明する。次に、この単語の意味表現 Word Space を文脈ベクトルに拡張した、InfoMap (Information Mapping)と呼ぶ情報検索システムの構成について述べる。また、InfoMap 情報検索システムにおいて分野依存の英語コーパス OHSUMED[Hersh 94]を用いた予備的な実験結果について報告する。

2． 連想情報検索と Word Space

2.1 単語の連想関係に基づく情報検索

大規模テキスト・データベース、例えば、図書館のカード・カタログや新聞記事の蓄積に対する全文検索では、単語の並びとしての質問(query)のどれかの単語、あるいはすべての単語を含む文書が返される。しかし、これらの質問中で使われた質問語(query word)を字面そのままではなく、それが表現する概念として扱った場合には、たとえその質問語が含まれていなかったとしても適切な文書を検索できると考えられる。我々の目標は知的な概念ベースの情報検索である。

我々の基本アプローチは Hinrich Schütze [Schütze 95] によって開発されたものである。まず、テキスト中の単語の共起の頻度、例えば、同じ文書中で近くに出現する2語の回数を記録することから始める。このとき、ある単語と内容表現語(content-bearing word)の集合との間の共起の分布は、単語の使われ方の輪郭(profile)、すなわち、単語の意味として役に立つと考えられる。異なった語の輪郭を比較することによって、それらの単語がいかに関連しているかの測度(類似度)を構築することができる。この単語の共起関係から導出された単語の類似度を一般化して、質問語の輪郭と各文書から生成された輪郭とを比較することによって、その文書のテキスト中にその単語自体が含まれていなかったとしても、概念的に質問語と関連していると判定した文書を検索することができる。

一般に、ベクトル空間モデル[Salton 75]を用いた検索方式を概念検索と呼ぶことがあるが、この手法を単語の連想関係(word association)に基礎を置くものとして連想情報検索(associative information retrieval)と呼んでいる。

2.2 特異値分解(SVD)による Word Space の構築

ある語と内容表現語の集合との間の共起頻度を高次元の空間をなす共起行列(co-occurrence matrix)に記録する。この抽象的な空間は類似した単語(より正確には、類似した分散的(distributional)な振る舞いを持つ語)が類似したベクトルを持つような概念空間である(表1)。

表１ 語彙の共起行列の例

単語＼内容表現語	…	market	…	last	…
Sunday		97	…	215	…
…	…	…	…	…	…
weekend		201	…	408	…
…	…	…	…	…	…

1000語

2万語

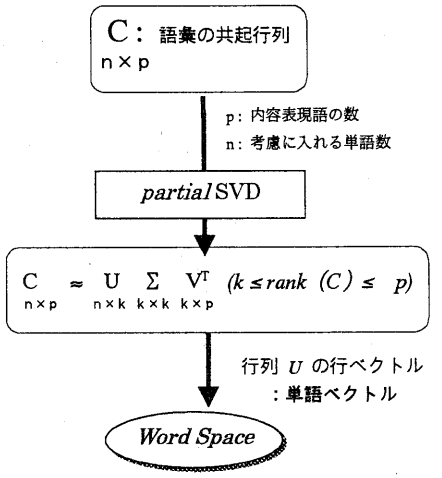

図１ Word Space のための部分 SVD

共起行列は２つの問題を抱えている。単語の特徴量(feature)が多すぎることと、データのスパース性である。これらの問題を解決するために、次元の縮退(dimensionality reduction)および一般化のツールとして、我々は共起行列に特異値分解（Singular Value Decomposition: SVD）[Golub 96] を適用する。SVD は任意の m × n の行列 A を以下のように分解(factorization)する線形代数の手法である。

$$A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V^T_{n \times n} \quad (1)$$

ここで、右辺の左の行列 U と右の行列 V は直交行列であり、固有ベクトルからなる行列である。真ん中の行列 Σ は対角行列であり、特異値行列と呼ばれる。また、添字 Tは転置を表わす。

式(1)は線形代数における完全(full)な SVD を示している。我々は式(1)を任意の大きさで近似した部分的(partial)な SVD の出力である左直交行列 Uを縮退行列として用いる(図１)。ここでは、この縮対行列の行を**単語**ベクトル(word vector)と呼ぶ。SVD は、ベクトルの特徴的な次元を取り出す計算手法であるため、この単語ベクトルが単語の意味の間の連想関係を近似していると考える。我々は、共起行列の概念空間を縮退して得たこの空間を Word Space と呼んでいる。この Word Space は２次の共起(second order co-occurrence)情報を含んでおり、この情報を通じて捉えられる単語の連想的(associative)な振る舞いを潜在的に反映していると考えられる。

LSI (Latent Semantic Indexing) [Deerwester 90] では、SVD を利用して語-対-文書の行列を縮退させ、情報検索に応用している。Word Space と LSI の違いは [Schütze 97] に議論されている。

単語ベクトルをその近接度(proximity)でクラスタリングすることによって、Word Space は単語の意味の曖昧性解消やシソーラス構築に用いることができる [Schütze 95,98]。

2.3 特異値分解(SVD)と主成分分析(PCA)

SVD は直接には統計手法ではなく線形代数における行列分解の手法である[Strang 93]。処理対象の行列の要素が統計的な観測値からなるとき、SVD は強力な統計解析のツールとなる。SVD は多変量解析における特徴量を削減する技法である主成分分析(principal component analysis : PCA)と密接な関係を持っている[Lay 97] [Schütze 95]。多変量解析は複数の変数（特徴量）の間の関連性(association)に関心があり、多変量データの輪郭の間の関係を発見することを目的としている。

今、行列 X が $p \times n$ のデータ行列(観測値行列)であると仮定しよう。行列 B がデータ行列 X の平均偏差形式(mean-deviation form)の行列であり、$A = (1/\sqrt{(n-1)})\ B^T$であるとき、

A^TA は不偏(unbiased)分散行列 S になる。このとき、固有値分解によって $p \times p$ の分散行列 S から p 個の固有値と固有ベクトルを計算することができる。

固有値分解は正方行列にのみ適用可能であるが、SVD は任意の矩形行列に適用可能である。したがって、SVD の計算は、固有値分解よりも便利が良い。

行列 A に SVD を適用したとき、Aの特異値(singular value)の２乗は、分散行列 Sの p個の固有値になり、行列 Aの右特異ベクトル $v_1 \cdots v_p$ は、行列 Xにおけるデータの主成分の係数となる。このとき、 $v_i^T X$ は第 i 番目の主成分である。従って、一般に SVD を主成分分析を実行するツールとして用いることができる。（図２参照)。

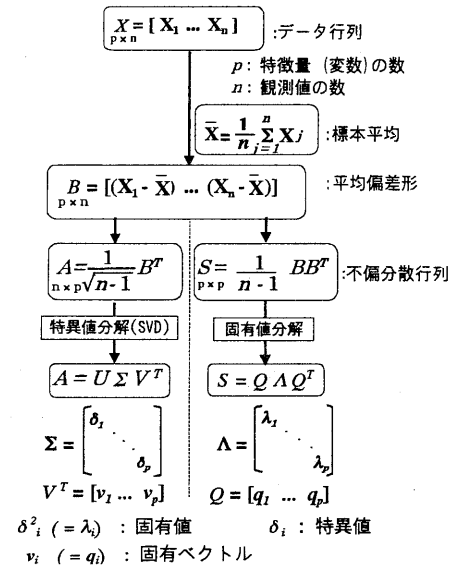

図２ SVD と PCA の間の関係

Word Space においては、係数 $1/\sqrt{(n-1)}$を持つ平均偏差形式の行列 A の代わりに、SVD を元のデータ行列 X(我々の場合には、図１の語彙共起行列 C)に直接的に適用している。

３．InfoMap 情報検索システムの構成

我々は、単語の意味表現である単語ベクトルを質問や文書の表現に拡張して、ベクトル空間モデルによる情報検索システム InfoMap を構築した。すなわち、文書と質問を単語と同様に高次元空間のベクトルとして表現する。

InfoMap は他の情報検索システムと同様に文書登録フェーズ（概念ベース Word Space を作成する）と文書検索フェーズからなる。この章ではこれらのフェーズの構成について説明する。

３．１ 語彙共起に基づく Word Space の構築

InfoMap の文書登録フェーズは、図３に示すようにコーパスにおける語彙の共起関係に基づく Word Space を構築する機能が中心である。

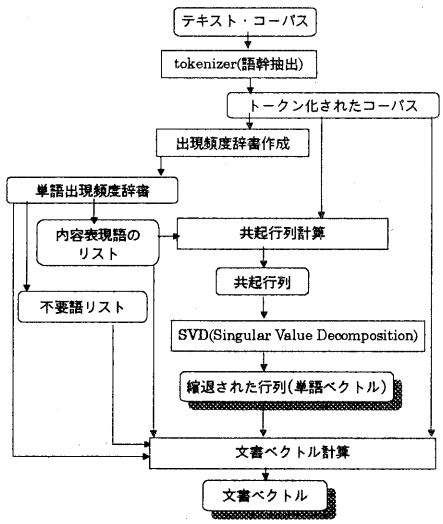

図３ Word Space の構築

3.1.1 テキストコーパスのトークン化

処理の最初の段階は、生のテキスト・コーパスから単語毎にトークン化されたコーパスを生成することである。英語コーパスに対する tokenizer による処理では、オプションとして語幹抽出(*stemming*)[Porter 80]を行なう。

3.1.2 単語の出現頻度の計算

処理の第２段階として、単語の出現頻度辞書(word count dictionary)を生成する。この出現頻度辞書は、コーパス中に出現する単語(トークン)とその頻度からなるリストであり、トークンの出現頻度の順に順位付けられる。

3.1.3 共起頻度の計算

コーパス中に出現する高頻度の2万語[1]に対して、1000次元の共起カウントのベクトル(共起行列)を生成する。これらのベクトルは各語の分布(distribution)の輪郭を表現する。各ベクトルの1000個の要素の値は、その単語の内容表現を決定する重みを表わす。

内容表現語は、単語のコーパス中での出現頻度、その単語のコーパス内の総出現頻度、その語の品詞情報、コーパスにおける文書内の単語の相対的な集中具合などを考慮し選択する。この計算は語の dispersion と呼ばれ、内容表現語としては文書全体を通じて均等に分布していない語が相応しいと考えられる。我々は、コーパス中で 51 番目から 1050 番目に頻度が高い 1000 語を内容表現語の基本集合として採用している。

内容表現語の一つの周りの特定の範囲内に2万語の一つが現れるたびに、ウィンドウの中に出現するそのベクトル中における適当な位置をカウントする。ウィンドウは、ある語が内容表現語からある距離の中にあるか、同一の文、パラグラフ、あるいは文書の中に入るかである。我々が用いているウィンドウの大きさの標準設定は、51(現在の注目単語と、その左25単語と右25単語)である。

コーパス全体を処理した後に、極端な数の効果をスムージングするために、各頻度カウントを変換(transformation)する。従って、実際の共起行列の (i,j)-番目の要素は以下の実数で表される。

$$c_{ij} = \phi(cooc_{ij}) \qquad (2)$$

ここで、 $cooc_{ij}$ は単語 i がコーパス全体を通して内容表現語 j からウィンドウ内に出現する共起頻度である。記号 ϕ は係数データの変換を表わす。我々は、平方根を基本変換として採用している。

[1] この章に示した次元の数は我々の実験の一つで用いたものである。この値は、システム構成のパラメータ設定で変更できる。

3.1.4 単語ベクトル(Word Space)の生成

2万個の共起ベクトル(共起行列の行)はそれぞれ 1000 次元の概念空間の点を表現している。概念空間を用いた計算をより扱いやすくするために、この空間の次元を低くする必要がある。共起行列の次元を縮退するために我々が用いたツールは SVD [Golub 96]である。

InfoMap システムにおける SVD の計算は、共起行列を SVDPack ソフトウェアパッケージ[Berry 92]に通すことで行なう。SVDPack は、高次元空間を一番低い次元で近似するために、最も重要な次元の特徴量を抽出する処理を繰り返し的に行う。

図1で部分SVDの出力として示した左側の直交行列 Uとして、p=100 次元に縮退した行列を作成する。正規化したベクトルを求めるために、縮退した行列 Uの各行をその長さで割って単位ベクトルに変換する。この正規化した縮退ベクトルを、語彙の共起関係から導出された Word Space の単語ベクトル u_i (i = 1, ..., 20,000) として扱う。

3.1.5 Word Space 上の文書ベクトルの生成

次に、各文書を 100 次元の文書ベクトル(document vector)に処理する。この処理は、コーパス中で最も出現頻度の高い2万単語に対してあらかじめ計算しておいた単語ベクトルを読み込み、文書中に出現するそれぞれの単語に対応する単語ベクトルを足しあわせることによって行う。

$$d_j = \sum_i w_{ij} u_i \qquad (3)$$

ここで、 d_j は、文書 j に対応する文書ベクトル、w_{ij} は文書 j における単語 i に対する重み、そして、u_i は文書 j 中に現れる単語 i に対する単語ベクトルである。重み w_{ij} の既定値は 1 とする。[Schütze 97]では、w_{ij} として tf・idf (term frequency・inverse document frequency)を用いている。

オプションとして、そのベクトルに対して情報として寄与しないと予期される一般的、共通的な単語を不要語(stop word)として無視しても良い。

我々は、コーパス中での頻度が1番目から50番目までの単語を不要語の基本集合として用いる。

コーパス中の各文書に対して、すべての文書ベクトルを計算した後、コーパス中での各文書の位置を指す文書索引とともに文書ベクトルをディスク上に保存する。

これらの文書ベクトルが占める 100 次元空間はコーパスから導出された文書の概念ベースを具体化し、これらのベクトルのそれぞれが、この空間内における文書の主題あるいは意味に対応する特定の位置を表現していると考えられる。この定式化は、概念空間に互いに近くにあるベクトルは、関連した主題を持つ文書に対応することを期待させる。

簡単のために、共起行列の次元が独立した単語のみからなるように説明してきたが、オプションとして、隣接した語の出現頻度の contingency table を基に χ^2-test を適用して抽出した統計的に意味のある句(statistical phrase) [Schütze 98] を利用することがある。この統計的な句とみなす単語対を求めるために、すべての隣接単語の頻度を計算し、それらの χ^2-値を計算する。この χ^2-値の上位のある数 (例えば、5000 対)の隣接単語を "sticky pair" と呼んで一単語とみなし、共起行列の行の要素として扱うことがある。

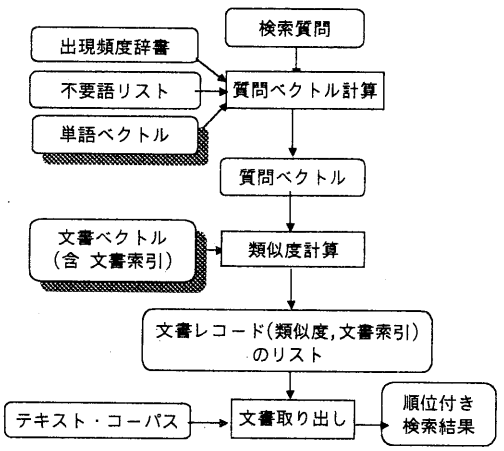

図4 Word Space 上での文書検索

3．2 Word Space 上での文書検索

InfoMap の文書検索フェーズの主な段階は、質問ベクトルの計算、質問ベクトルと文書ベクトルとの近接度（類似度）の計算と実際の文書の取り出しである(図4)。

3.2.1 質問ベクトルの計算

Word Space における連想関係を用いた文書検索を行なうために、まず、単語のリストの形で与えられた検索質問中の各単語に対応する正規化単語ベクトルの集合を取り出し、それらを足しあわせることで**質問ベクトル**（query vector）を形成する。

$$q = \sum_i w_i u_i \qquad (4)$$

ここで、q は質問ベクトル、w_i は単語 i の質問における重み（規定値 1）、u_i は質問中の単語 i に対応する単語ベクトルである。

3.2.2 文脈ベクトルの近接度の計算

上記の質問ベクトルを各文書ベクトルと比較し、質問ベクトルと最も近い文書ベクトルを持つ文書を検索結果として返す。

質問ベクトルや文書ベクトルは正規化された単語ベクトルの和（セントロイド centroid)である。これらのベクトルを一般に**文脈ベクトル**(context vector)と呼ぶ。

2つの文脈ベクトル（質問ベクトル q と文書ベクトル d_j）の近接度(*closeness*)は、このベクトル間の角度のコサイン計算によって決定する[2]。

$$\text{closeness}(q, d_j) = (q \cdot d_j) / (|q| \cdot |d_j|) \qquad (5)$$

類似度計算ルーチンは文書ベクトル(d_j)を入力とし、質問ベクトル(q)との近接度（類似度）によって順位づけられた文書レコードのリストを返す。各文書レコードはその文書の近接度の値と文書索引（コーパス中での文書の位置）を含む。

文書取り出しルーチンは単に文書レコードにに格納されている文書索引で検索し、その検索した文書をユーザが要求した文書として表示する。

[2] 類似単語の近接性を計算する場合もコサイン測度を用いる。

4．領域依存のコーパスを用いた検索実験

これまでの節で述べてきたように、Word Space における連想関係は、注釈のない(unannotated)テキスト・コーパスから外部知識なしに unsupervised な方法で計算することができる。我々は、この単語の連想関係(word association)が情報検索に有用であることを示したいと考えている。

これまで InfoMap システムにおいては、主に新聞記事コーパスをソースとして一般的(general)な単語の連想関係を取り出し、検索対象としても新聞記事を用いて実験を行なってきた[Schütze 97]。我々は、現在、異なった種類のトレーニングコーパスから抽出した単語の連想関係が検索処理の振る舞いにどのような結果を与えるかに関心を持っている。そこで、対象領域に依存したコーパスから抽出した領域依存の単語の連想関係(domain-specific word association)に基づく情報検索の実験を行なうことにした。

領域依存のコーパスの例として、医学文献のコーパス集である MEDLINE の一部(1987 年から 91 年までの５年分)を情報検索の評価用に編集した、OHSUMED (348,566 文書)と呼ばれるコーパス[Hersh 94]を用いた実験を進めている。OHSUMED には評価用の106個の検索質問文と各質問文に関連する文書の情報（専門家によって判定されたもの）が添付されている。

OHSUMED の文書データには、"definitely relevant" (DR：検索質問文に完全に適合する文書）と "possibly relevant" (D＋PR：検索質問文におそらく適合する文書）の２種類の解答が添付されている。InfoMap システムでは、潜在的に関連のある文書（D＋PR）の検索に効果を発揮することを狙っている。残念ながら、[Hersh 94]には DR の結果しか載っておらず、D＋PR の結果については直接比較することができない。そこで、まずに DR に関して評価してみることにした。表２に実験結果の一部を示す。

[Hersh 94]に記載されている SMART システムでの５つの手法（パラメータの設定などを変えて５種類の実験結果が記載）による OHSUMED に対する再現率(recall)の平均は、検索要求との類似度順にランク付きで出力された上位５文書、15 文書、100 文書に対して各 11.5、21.7、48.8(%)である。

表２ OHSUMED を用いた検索実験の結果(再現率)

		再現率(recall) %		
	文書数	5 文書	15 文書	100 文書
Smart 平均	348,566	11.5	21.7	48.8
InfoMap	文書数	5 文書	15 文書	100 文書
(20k 語)	54,710	12.5	23.2	51.5
(30k 語)	54,710	13.8	25.0	53.3
(30k 語)stem	54,710	14.6	23.8	54.2
(30k 語)	124,535	7.39	17.4	38.7
(30k 語)stem	124,535	9.24	16.6	41.8

・() 内は共起計算に用いた単語数。
・内容表現語はいずれも 1000 語。
・stem は、語幹抽出をしたことを表す。

1987 年分の約５万４千文書に対して評価した InfoMap での実験結果は、再現率 12.5、23.2、51.5(%)であり、比較的良い結果を得た。しかし、1987-88 年分の約12万４千文書では再現率がかなり低くなってしまうことがわかった。以下、文書数が増えると次第に再現率が低くなっていく（表２では３年分以上のデータに関する結果の記載を省略）。今回の実験では、実験対象とした文書全体から共起計算を行なって作成した単語ベクトルを用いて文書ベクトルを生成した(close テスト)ため、過剰学習により性能が劣化したものと思われる。

OHSUMED コーパス全体から辞書として抽出した異なり単語見出しの総数は 27.3 万語である。(語幹抽出処理を施した場合は、22.8 万語）。新聞記事を対象とした従来の InfoMap の検索実験では、共起行列の大きさとして２万語×1000 語のものを用いてきたが、OHSUMED では上記のように文書中に出現する異なり語数が多いので、３万語×1000 語の共起行列を用いたところ、表２に示すように再現率が改善された。

表2で、stemと示した行は語幹抽出(stemming)ありの場合、他はなしの場合である。上位5文書および上位100文書の場合に語幹抽出の効果が現れている（上位15文書の場合には語幹抽出をした方が再現率が低くなっている）。

この実験では、3.1.5節で述べた統計的に有意な句を一単語とみなして共起行列に加える処理は行なっていない。統計的な句を考慮することにより、検索効率の改善が期待できると考えている。

5．まとめ

情報検索の対象となるテキスト文書は単語の集まりから構成されており、単語の意味表現は情報検索システムの基礎として重要である。本研究では、コーパスにおける単語の共起関係から導出したベクトル空間 Word Space における単語ベクトルによって単語の意味を表現した。単語ベクトルを情報検索等に利用する場合には、ベクトルの次元を小さくしておく必要がある。我々は単語ベクトルの次元を縮退させるために線形代数の手法である特異値分解(SVD)を用いた。本稿では、Word Space における SVD の利用が主成分分析(PCA)とほぼ等価な計算を行なっていることを説明した。

また、単語ベクトルを文書ベクトルや質問ベクトルといった文脈ベクトルに拡張し、情報検索システム InfoMap を構築した。この報告では、InfoMap 情報検索システム構成の構成と、InfoMapによる領域依存の英語コーパスを用いた予備的な実験結果を述べた。引き続き、領域依存のコーパスが検索処理の振る舞いにどのような結果を与えるかを検討していく予定である。

関連研究として、電子メールを用いて個人の興味を反映するようにした個人的(personal)な単語の連想関係をベースにした実験を行なっている[Flournoy 98]。また、InfoMapの手法が多言語情報検索に適用できるかを検討するために、英語-日本語間の自立語の用語リストの翻訳に適用する実験[Kikui 98]を行なっている。

謝辞

本研究は第一著者がスタンフォード大学滞在中に行なったのものである。Dr.Hinrich Schütze（Xerox Palo Alto研究所）をはじめ本研究に協力いただいたスタンフォード大学 InfoMap 研究チームのメンバー各氏、および滞米研究に対して支援いただいた三菱電機(株)の諸氏に感謝します。

参考文献

[Berry 92] Michael W. Berry: *Large Scale Singular Value Computations*, International Journal of Supercomputer Applications, 6:1, pp. 13-49, 1992.

[Deerwester 90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman: *Indexing by latent semantic analysis.* J. American Society for Information Science, 41(6):391-407, 1990.

[Flournoy 98] Raymond S. Flournoy, Ryan Ginstrom, Kenichi Imai, Stefan Kaufmann, Genichiro Kikui, Stanley Peters, Hinrich Schütze, Yasuhiro Takayama: *Personalization and Users' Semantic Expectations* , ACM SIGIR'98 Post-Conference Workshop on Query Input and User Expectations, 1998.

[Golub 96] Gene H. Golub, Charles F. Van Loan: *Matrix Computation*, 3rd ed., The Johns Hopkins University Press, 1996.

[Hersh 94] W. R. Hersh, C. Buckley, T. J. Leone, D. H. Hickam: *OHSUMED: An interactive retrieval evaluation and new large test collection for research*, Proc. 17th Annual ACM SIGIR Conference '94, pp. 192-201, 1994.

[Kikui 98] Genichiro Kikui: *Term-list Translation using Mono-lingual Word Co-occurrence Vectors*, Project Note, COLING-ACL '98, 1998.

[小嶋 97] 小嶋秀樹, 伊藤昭:"文脈依存的に単語間の意味距離を計算する一手法",情報処理学会論文誌, Vol.38, No.3 1997.

[Lay 97] David C. Lay: *Linear Algebra and its applications*, revised ed., 1997.

[丹羽 94] 丹羽芳樹, 新田義彦:"単語ベクトルを用いた多義語の意味推定—共起ベクトルと定義距離ベクトルの比較—",情報処理学会自然言語処理研究会,102-7,1994.

[Porter 80] M. F. Porter: *An algorithm for suffix stripping*, Program, 14, pp.130-137, 1980.

[Salton 75] Gerard Salton, A. Wang, C. S. Yang: *A vector space model for automatic indexing*, Comm. ACM, 18, pp.613-620, 1975.

[Schütze 95] Hinrich Schütze: *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*, CSLI Lecture Notes 71, CSLI Publications, 1997.(Ph.D. thesis, Stanford Univ., Dept. of Linguistics, July 1995.)

[Schütze 97] Hinrich Schütze, Jan O. Pedersen: *A cooccurrence-based thesaurus and two applications to information retrieval*, Information Processing & Management, Vol.33, No.3, pp.307-318, 1997.

[Schütze 98] Hinrich Schütze: *Automatic Word Sense Discrimination.* Computational Linguistics, Vol. 24, #1, pp.97-123, March 1998.

[Strang 93] Gilbert Strang: *Introduction to Linear Algebra*, Wellesley-Cambridge Press, 1993.