自然言語処理　144－8
（2001．7．16）

統計的モデルを用いた単語クラスタリング

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
TEL:(03)5452-5277　FAX:(03)5452-5278¥¥
{kawamae, aoki, yasuda}@mpeg.rcast.u-tokyo.ac.jp

既存の検索システムはユーザのニーズそのものでなく、キーワードによって検索を行っている。検索にキーワードを利用するため、自然言語の持つ多様性によってユーザの検索が非効率になることがある。この問題の解決の一つにシソーラスがある。そこで本稿では分野と目的を特化したシソーラスの構築を目的として、統計的な単語クラスタリングの手法を提案する。具体的には文書集合から文書の背後にある概念を抽出し、それぞれの概念に固有な単語をクラスタリングする。同一概念毎に単語がクラスタリングされることで分野が特化し、検索質問の拡張あるいは検索結果の構造化といった目的に対応したシソーラスの自動構築が可能となる。提案手法を学術的な内容の文書集合に適用して、単語クラスタリングを生成し、概念毎の単語クラスタリングが生成された結果を報告する。

情報検索　シソーラス　概念検索　単語クラスタリング　特異値分解　因子分析

The Word Clustering Based on Statistical Model

Noriaki Kawamae, Terumasa Aoki, Hiroshi Yasuda
Research Center for Advanced Research and Technology, The University of Tokyo
4-6-1, Komaba, Meguroku, Tokyo, 153-8904, JAPAN
TEL:+81-3-5452-5277　FAX:+81-3-5452-5278
{kawamae, aoki, yasuda}@mpeg.rcast.u-tokyo.ac.jp

The existing search systems are based on simple word matching method. Therefore the variety of natural language prevent user search activity. The thesaurus is one answer to this problem. We propose a novel statistical word clustering to construct the thesaurus automatically. Here, the concepts are extracted from documents and words in documents are clustering into the same concepts. We can construct the thesaurus that is specialized on a domain and in a function by the word clustering. The proposed method is applied to a set of conference documents to examine the effectiveness of the generated word clustering.

Information Retrieval, Thesaurus, Conceptual Search, Word Classification, Factor Analysis

1. はじめに

我々がインターネットやデータベースなどから情報を検索するシステムはキーワード検索である。キーワード検索は、検索システムにユーザがキーワードを入力すると、システムは検索結果としてそれらのキーワードを含む文書を表示する。キーワードさえ入力すれば本の索引からそのキーワードを含む個所を探すのと同じである。インターネット及びデータベースから入手可能な電子化された情報が増加するのに伴い、効率的な検索の実現のために、ユーザの検索システムへのニーズは高まるばかりである。

このキーワード検索は問題点が二つある。第一に、キーワードが想起できなければ検索が開始できないことである。その理由は、検索システムは本と異なり索引がない。ユーザは自身の検索ニーズそのもので検索しているのではなく、現状は検索ニーズにあった文書に含まれていそうな単語を想起し、それをキーワードとして検索している。従って、検索ニーズを的確に表したキーワードを想起する必要がある。第二に、キーワード自身の問題である。それは言語を利用することによって生じる語彙不一致と多義語の存在である。語彙不一致は同じ概念を現すのに、文書作成者によって利用する単語が異なることである。単語が多義語であれば、同じ単語で違った概念をあらわす場合がある。従って、検索対象となる文書集合にユーザの検索ニーズを満たすものがあっても、想起した単語でそれらの文書を検索できない問題がある。

ユーザが効率的な検索を実現するためにはこれらの問題を解決する検索支援手法が必要となる。その問題を解決するのにシソーラスの構築が考えられる。シソーラスは類義語、同義語などを含んだ意味や概念の辞書である。シソーラスの構築は人手を要するので非常に高価で汎用性のあるものの構築は困難とされている。本研究は目的と分野を特化したシソーラスの自動構築を目標とする。本稿はその構築に必要な統計的な単語のクラスタリングについて提案する。

本稿で提案する単語のクラスタリングは統計モデルを導入する。文書中に出現する単語もまた、ユーザが入力したキーワード同様に背後に概念を持つと仮定する。この概念は文書作成者の文書作成者の意図である。導入する統計モデルは文書に出現した単語集合からこれらの概念を抽出するものである。同概念から出現した単語をクラスタリングすることで、目的と分野を特化したシソーラスの構築が可能と考えられる。

本研究は、文書に出現した単語集合から、その発生源となった概念を推定し、同概念の単語をクラスタリングする手法を提案した。実際の文書集合に対して提案手法を適用したところ、同概念における単語のクラスタリングが実現できることが確認できた。論文の構成は以下のようになっている。2 章では既存研究を振り返り、3 章では単語クラスタリングに必要な単語・文書行列と統計モデルを提案する。4 章では実験を行い、5 章でまとめる。

2. 既存研究

キーワードによる検索支援にシソーラスがある。シソーラスは語と語との関係を上位／下位部分／全体、類似、反意などの関係で分類・整理した辞書であり、語彙不一致の解消に利用されることが多い。しかし、その構築は非常に高価で、www ページの検索を支援する汎用のシソーラスの構築は困難である。また分野に依存しない汎用のシソーラスの利用は検索効率が低下することが多いと報告されている[11]。シソーラスの構築が困難な原因は、人手によって作成するためである。これを自動化すれば問題し、支援する分野を特定すれば、検索効率を支援できるシソーラスの構築の可能性がある。

そこで本稿は特定分野のシソーラスの自動構築に利用可能な単語クラスタリング手法を提案する。単語のクラスタリングは、最適化基準を用いて単語全体の集合をいくつかのグループ（クラスタ）に分類することである。単語のクラスタリングは単語のどの属性に着目するか、属性値間の類似度、クラスタリングアルゴリズムによって様々な方法が考えられる。代表的なクラスタリングアルゴリズムには次のものがある。Bellegarda[1] は単語・文書行列を用いてクラスタリングを行っている。Schutze[6],[7], Schutze & Pedersen[8] は単語・単語行列を用いている。Hughes & Atwell[2], Hughes[3] は単語の文中における位置情報によって行い、Hogenhout & Matsumoto[4] は構文解析済みコパースから抽出された単語間の係り受け関係を用いている。これらの手法は特異値分解 (Singular Value Decompostion;SVD) に基づいた手法である[9]。この手法は文書中に隠された単語の意味、概念的な相関を抽出するために用いられる。手法の概要は、観測されたデータのベクトル空間を、できるだけ情報の損失を小さく、SVD が用いて行うものである。文書集合の分類に SVD を用いると、出現回数の多い共起する単語の組を共有する文書は、潜在的意味空間において類似関係を持つことになる。また、単語のノイズに対して頑健な特徴も持つ。この手法は文書の分類だけでな

く、単語のクラスタリングにも適用できる。

SVDを用いた文書分類は、SVDにより導出された潜在的な意味空間での類似度を用いることで、単語のノイズの影響に対して頑健な分類・検索ができるが、モデルの存在を仮定していない。また、潜在的な意味空間を構成する軸は、出現した単語から合成された軸である。これは文書間の関係を観測された単語の軸で表現し、その単語発生の原因となる概念は利用されていない。提案手法は単語の出現の背後に概念の存在を仮定している。概念を反映したモデルを利用することでより精度の高い単語のクラスタリングが期待できる。

3. 提案手法

本稿は、同じ概念から出現した単語をクラスタリングする手法を提案する。提案手法は、次の手順から構成される。

(1) 単語・文書行列の作成：各文書からのクラスタリングに有効な単語集合を抽出し、その重みを計算し単語・文書行列を作成する。

(2) 概念の抽出：単語・文書行列から概念空間を求める。

(3) クラスタリング：概念空間における単語をクラスタリングする。

1. に関しては3.1で、2. に関しては3.2で詳細を述べる。

3.1 単語・文書行列の作成

単語の属性値を決定すれば、各文書はそれを要素とする文書ベクトルで、文書集合は単語・文書行列の形式で表現できる[9]。単語・文書行列Aは単語を行、文書を列とする行列である。単語をw_i、文書をd_j、とすると、行列の要素(i,j)はその単語の重みa_{ij}となる。単語w_iはこの重みを成分とするベクトルで表現でき、ベクトル空間に配置することができる。重みは大局的重み付けと局所的重み付けの二通りがあるが、これらを組み合わせて利用する。下に挙げるL1はtf(Term Frequency)、G3はidf(inverse document frequency)として呼ばれ、よく利用される。

局所的重み付けは文書d_j内の単語に対してのみ重み付けを行う

L1:出現頻度

$$P_{ij} = \frac{C_{ij}}{C_j} \tag{1}$$

P_{ij}:文書d_jにおける単語w_{ij}の出現頻度

C_{ij}:文書d_jに出現した単語w_iの数

C_j:文書d_jに出現した単語の数

L2:正規化されたエントロピー

$$H_{ij} = -\frac{1}{\log M}\sum_{i=1}^{M} P_{ij}\log P_{ij} \tag{2}$$

M:文書d_jに出現した単語の数

エントロピーの定義よりH_{ij}の取りうる値は$0 \leqq H_{ij} \leqq 1$となる。各単語が等確率で出現するほど1に近く、限られた単語しか出現しない場合は0になる。重み付けとして利用する場合、次の変形を行う。

$$G_{ij} = 1 - H_{ij} \tag{3}$$

大局的重み付けは文書集合全体に渡って重み付けを行う。

G1:文書全体における頻度

$$P_i = \frac{C_i}{C} \tag{4}$$

P_i:文書全体における単語w_{ij}の出現頻度

C_i:文書集合に出現した単語w_iの数

C:文書集合に出現した単語の総数

G2:単語毎のエントロピー各単語の文書全体における出現頻度

$$H_i = -\frac{1}{\log N}\sum_{j=1}^{N} P_{ij}\log P_{ij} \tag{5}$$

N:文書集合に含まれる文書の総数

P_{ij}:文書d_jにおける単語w_{ij}の相対頻度

L2が文書内での単語の正規化されたエントロピーであったのに対し、G2は文書全体における単語の正規化されたエントロピーとなる。同様に重み付けとして利用する場合、次の変形を行う。

$$G_i = 1 - H_i \tag{6}$$

G3：文書数の逆数

$$H_i = -\log\frac{C_d}{C_{w_i}} \tag{7}$$

C_d:文書集合に含まれる文書数

C_{w_i}:単語w_iを含む文書の数

3.2 概念の抽出

単語発生の概念を推測する為にモデルを設定する。このモデルは、「文書に出現した単語はその背後に単語発生の原因となる概念を持つ」を仮定する。概念が推測されることで単語はベクトル空間から概念空間へ変換される。また空間を構成する概念の軸の数も文書の軸の数よりも少ないことが求められる。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。因子分析を用いて定式化すると次のようになる。

$$a_{ij} = w_{i1}c_{1j} + w_{i2}c_{1j} + \ldots + w_{in}c_{nj} + u_i v_i \tag{8}$$

a_{ij}:単語w_iの文書d_jにおける観測値

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 における概念 c_m の得点
w_{im}:因子負荷量。単語 w_i と概念 c_m の相関
v_i:独自因子得点。単語 w_i に固有な得点
u_i:独自因子負荷量。単語 w_i と独自因子得点 v_i の相関
$m \leqq j$:概念の個数は文書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

$$A = WC + VU \tag{9}$$

W : 共通因子パターン行列、$(i \times m)$ 型行列
C : 共通因子行列、$(m \times j)$ 型行列
U : 独自因子パターン行列、$(i \times i)$ 型行列。対角成分の i 番目が文書 di の独自因子負荷量、他の成分は 0
V : 独自因子得点行列、$(i \times j)$ 型行列
ここで観測されたデータは a_{ij} のみである。因子分析の目的は文書 dj について独自の部分を出来るだけ小さく、因子負荷量を推定することになる。この推定には独自部分の評価と概念の個数をあらかじめ決めておく必要がある。従来は個々を別々に決定していたが、確率的コンプレキシティ (Stochastic Complexity:SC)[5]、[12] を用いて同時に決定する。SC はモデルのパラメターを m として固定して符号化した場合、最も短く符号ができる場合の符号長という意味を持つ。ここで概念の個数の決定に用いた SC は次のように式になる。

$$SC\{A \mid m\} \cong -\log P_{\hat{a}}\{A\} + \frac{m}{2}\log n \tag{10}$$

$\log P_{\hat{a}}\{A\}$: 単語・文書行列 A の最尤推定量 m : 概念の個数
n : 単語・文書行列 A の要素数
n : 単語・文書行列 A の要素数

3.3 特異値分解

行列 A の特異値分解のモデルは次のように定義される。

$$A = U\Sigma V^T \tag{11}$$

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ \vdots & \vdots\vdots & & \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}$$ 単語・文書行列

$(n \times m)$ 型行列

$UU^T = V^TV = I$(単位行列)

U は $(n \times r)$ 型行列 V は $(r \times m)$ 型行列)

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots\vdots & & \\ 0 & 0 & \dots & \sigma_r \end{pmatrix}$$ 対角行列

$(r \times r)$ 型行列

$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$

σ_k は行列 A の特異値と呼び、AA^T の固有値の非負の平方根に等しい。ここで大きい順に m 個の特異値 $\sigma_1 \ldots \sigma_m$ だけを用いて再構成された行列 A_k は次のようにして求められる。

$$A_k = U_k\Sigma_k{V_k}^T \tag{12}$$

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ことがある。モデル選択の指標は SC を適用する。

4. 評価実験

提案した単語クラスタリングの手法ついて、次の二点で評価する。一つは概念の推測の可能性である。本稿で提案する単語クラスタリングは文書集合に出現した単語からその概念を推測し、クラスタリングを行うことである。そこで提案手法によって概念が抽出できるかを評価する。次にクラスタリング結果を評価する。実験に利用した文書は全部で 60。文書の内訳は情報理論、情報検索、統計学について解説されたものをそれぞれ 20 用意した。これらの文書を形態素解析を行い、品詞毎に分類する。形態素解析には[10] を用いた。今回の実験で用いた単語は名詞と未知語である。その理由は検索においてユーザが利用するキーワードの大半がこれらに該当するからである。未知語とは形態素解析に用いた辞書に登録されていない単語である。また、三つ以上の文書に渡って出現しない単語は解析の対象外とした。大きさが 1 の単語はストップワードとして解析の対象から外す事が多い。今回の実験ではストップワードを英数字の一文字だけのものにした。日本語の場合は漢字一文字でもそれなりの意味を持つのでストップワードに入れなかった。利用した文書集合に含まれる単語の異なり数は 4692 であったが、この制約を入れると 522 となった。

4.1 概念の推測

表 1 は様々な重みを持つ単語・文書行列から抽出された概念空間の固有値が最も高い軸において因子負荷量が高い順に単語を 10 個並べたものである。その下は単語の各カテゴリにおける割合を示したものである。いずれも統計学で高い値を示しており、これらの単語からこの軸は統計学に関しての概念であると考えられる。重みとしては *L2G2*、*L2G3* 共に高い精度を持っていた。この重みを持つ単語・文書行列を主成分分析に利用した結果も併せて示しておく。

表 2 と表 3 を比較すると、因子分析ではそれぞれの

表 1 単語・文書行列と手法の違いによる単語クラスタリング
Table 1 Word Clustering based on different method

手法	主成分分析		因子分分析					
重み付け	L2G2	L2G3	L1G1	L1G2	L1G3	L2G1	L2G2	L2G3
単語	結果	差	単純	単純	単純	単純	単純	単純
	データ	確定	特殊	任意	任意	これら	標本	標本
	例	散布	これら	下記	下記	特殊	項	項
	平均	得点	誤差	σ	σ	適切	変数	変数
	簡単	現象	適当	項	項	様々	推定	推定
	問題	主成分	条件	母	母	多次元	共	共
	度	近似	多次元	他方	他方	条件	母	一定
	推定	理解	最小	未知	期待	行列	別	誤差
	標本	数値	モデル	未知	期待	行列	別	誤差
	統計	λ	λ	期待	未知	モデル	誤差	仮定
情報理論	15.6	2.6	5.7	1.7	1.5	5.8	1.7	0.8
情報検索	16.4	0	7.2	1.2	1.4	7.8	0.0	0.0
統計学	67.9	97.3	87.2	95.3	97.0	86.3	98.2	99.1

表 2 因子分析による単語クラスタリング
Table 2 Word Clustering based on factor analysis

カテゴリ	情報理論		情報検索		統計学	
評価指標	再現率	適合率	再現率	適合率	再現率	適合率
＋＋＋	3.0	8.23	1.9	4.9	29.5	86.8
＋－－	5.5	21.4	20	77.0	0.45	1.6
＋＋－	1.6	4.2	1.7	5.1	31.3	90.6
＋－＋	8.3	16.8	11.5	30.1	22.5	51.7
－＋＋	17.5	76.3	3.8	18.8	1.3	5.0
－－－	2.5	11.3	15.0	71.3	3.8	17.5
－＋－	5.0	25.0	12.5	62.5	2.5	12.5
－－＋	9.5	34.2	13.4	51.4	4.8	14.4

＋、－ は左から固有値の高い軸における因子付加量の符号を表している

表 3 主成分分析による単語クラスタリング
Table 3 Word Clustering based on principal analysis

カテゴリ	情報理論		情報検索		統計学	
評価指標	再現率	適合率	再現率	適合率	再現率	適合率
＋＋＋	4.5	12.6	8.4	27.4	22.5	60.0
＋－－	4.9	14.4	11.4	34.4	21.6	51.2
＋＋－	6.4	15.7	6.3	17.4	25.6	65.3
＋－＋	8.0	21.4	0.8	28.8	22.2	49.7
－＋＋	3.0	9.2	3.6	8.0	26.9	82.8
－－－	3.2	8.8	4.2	13.4	24.3	76.3
－＋－	3.7	9.4	5.0	14.2	25.7	76.3
－－＋	4.7	12.6	4.0	10.8	27.1	77.1

カテゴリの適合率・再現率の高いクラスタが存在するが、主成分では殆ど統計学のカテゴリに偏っている。その統計学関連用語のクラスタでも再現率・適合率が因子分析によるものよりも高いクラスタは存在しなかった。因子分析、主成分共に統計学のクラスタが多かったが、各カテゴリで再現率、適合率がそれぞれ最も高いクラスタが存在する因子分析による手法が同概念の単語をクラスタリングするのに適切であることが確認された。主成分分析と因子分析を比較の違いはモデルの有無である。このモデルの有無が概念の抽出に影響を与えると考えられる。

4.2 クラスタリング

次に単語のクラスタリングを *L2G3* から抽出された概念空間においてクラスタリングを行った。手法は K-Means による。グループは表 2 と対応している。

表 2 、表 4 より各カテゴリに特徴的な単語のクラスタリングが確認された。同クラスタ内部の単語の関係は類義語、同義語より同一文書に共起しやすいものと考えられる。

5. ま と め

本稿は単語のクラスタリングを、単語の概念に基づ

表 4 単語クラスタリング
Table 4 Word Clustering Results

グループ	＋＋＋	＋－－	＋＋－	＋－＋	－＋＋	－－－	－＋－	－－＋
単語	同時	重	最適	上	符号	文書	アルゴリズム	語
	密度	群	水準	中	圧縮	再現	我々	サービス
	積分	因子	違い	簡単	訂正	報告		インターネット
	微分	グラフ	最小	関	効率	学会		通信
	∂	偏差	検討	直接				データベース
	前記	主	個体	的				画像
	σ	解釈	連立	方				発展
	無関係	主成分	個数	部				月
	右辺	共通	α	論				権
	μ	得点	否	日				今後

単語数が 10 未満のグループはカテゴリに含まれる全ての単語である

いて行う手法を提案した。提案手法は文書集合の背後にある概念を推測する。この手法によって文書は文書を軸とするベクトル空間から推測された概念を軸とする概念空間に配置される。これを実験に適用した結果、従来の主成分分析よりもクラスタリング結果は同概念からの単語が含まれる結果を得た。その理由は概念空間の類似関係は単語を軸としたベクトル空間は SVD による空間よりも単語の本質的な内容の類似性を反映していると考えられる。最後にシソーラースへの適用としては類義語・同義語は少ない。これらの単語を組み合わせて同概念を現すので、検索質問拡張や含んでいる単語で検索結果の分類といった目的に寄与できると考える。

参 考 文 献

1) Bellegarda, J. R., Butzberger, J. W., Chow, Y-L., Coccaro, N. B & Naik, D.: A novel word clustering algorithm based on latent semantic analysia. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96), 1996.
2) Hughes, J. & Atwell, E.: The automated evaluation of inferred word classifications. In Processings of European Conference on Artificial Intelligence 1994.
3) Hughes, J.: Automatically Acquiring a Classification of Words. Ph.D Thesis, School of Computer Studies, The University of Leeds 1994.
4) Hogenhout, W. R. & Matsumoto, Y.: A preliminary study of word clustering based on syntactic behavior. In Processings of the 1997 Meeting of the ACL Special Interest Group in Natural Language Learning 1997.
5) J. Rissanen.: Fisher information and stochastic complexity. IEEE Transactions on Information Theory, 42(1):40–47, January 1996.
6) Schutze, H.: Word sense disambiguation with sublexical representaions. In AAAI-92 Workshop Notes, Statistically-Based NLP Techniques 1992.
7) Schutze, H.: Part-of-Speech induction from scratch. In Processings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93) 1993.
8) Schutze, H. & Pedersen, J.: A vector model for syntagmatic and paradigmatic relatedness. In Processings of the 9th Annual Conference of the University of Waterloo Centre for the New OED and Text Research 1993.
9) 北研二: 確率的言語モデル, 東京大学出版会,1999.
10) 茶筅:http://chasen.aistnara.ac.jp/index.html.ja
11) 徳永健伸: 情報検索と言語処理, 東京大学出版会,1999.
12) 李航, 山西健司.: 線形結合モデルを用いた統計的語彙的トピック分析,IBIS2000.