データベースシステム 125－43
（２００１．７．１８）

情報理論的モデルを用いた情報検索

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
{kawamae, aoki, yasuda}@mpeg.rcast.u-tokyo.ac.jp

従来の検索システムによる文書の検索はキーワード検索で行われている。この手法による検索は自然言語を用いることと、利用者の検索へのニーズそのもので検索を行っていないために、ユーザが必要な文書を探し出すのに手間がかかる。そこで本稿は文書の検索を、文書の概念に基づいて行う手法を提案する。提案手法は文書に出現した単語でなく、単語からその発生源となる文書の概念・内容を推測する。その結果、文書が従来の単語空間から概念空間に配置される。概念空間における文書の位置関係が、単語空間よりも文書の内容に基づいた類似関係を反映するために、従来の手法よりも検索の精度が向上した。この手法を用いることでユーザは概念に基づいた検索が出来るようになる。

情報検索，情報理論，概念検索，確率的コンプレキシティ，文書分類，　潜在意味空間

The Information Retrieval Based on the Information Theory 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
{kawamae, aoki, yasuda}@mpeg.rcast.u-tokyo.ac.jp

This paper presents a novel a novel approach mapping documents into a conceptual space. Many search systems are based on simple word matching method. We have trouble in seeking the information by this method. Because this search are used by natural language and not by user's needs to information. Our presented information retrieval method use not words but concepts generating words in documents. This method is based on the information theory. We infer the concepts from words and map documents in the concept space. The relation of documents mapped in the concept space approximates the essential similarity between documents. Therefore the precision of document classification improves, and users can search by their concepts.

Information Retrieval, Information Theory, Conceptual Search, Stochastic Complexity, Document Classification, Latent Semantic Space

1 はじめに

ネットワークの利用者の増加やデータベースの普及に伴って、そこから我々が入手可能な電子化された文書の数も増大している。我々は電子化された必要な情報をこれらの中から見つけるために検索システムを用いる。現在の検索システムはキーワード検索を採用している。この検索方法は、検索システムにユーザがキーワードを入力すると、検索システムは検索結果として、そのキーワードを含む文書を表示する。キーワード検索は本での検索と違い、索引がないので自身でキーワードを入力しなければならない。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、キーワードが想起できなければ検索が開始できない。しかし、一旦、キーワードを検索システムに与えてしまえば、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。故に、ユーザは必要な情報を入手するために自身の検索ニーズをキーワードとして正確に具体化する必要がある。検索対象となる文書もまた単語の組み合わせで構成されている。キーワードも単語であるが、本稿ではユーザ側が用いたということで区別してこう呼んでいる。

キーワード検索は、ユーザの検索へのニーズそのものでなく、自然言語を用いて行われるので、次のような問題がある。まず一つは単語のゆらぎである。言語の利用には個人差があり、同じ内容の文書でも利用される単語は異なってくる。表記の違いもこの問題に含まれる。従って、ユーザの検索へのニーズと文書の内容が一致しても、キーワードと単語が一致できないので検索できないことがある。検索システムは内容でなく単語で検索するので、出現した単語だけでは内容の類似性に基づいた検索よりは類似した単語集合の検索になっている。次に単語の多義性の問題がある。多義語は利用される分野や文脈により意味は異なってくる。従来のように個々の単語で処理するのではなく、他の単語を利用して曖昧性を解消するなどのアプローチが必要である。

以上からユーザが文書を効率的に検索する為には、文書に含まれる単語の有無でなく、内容の類似性によって文書を検索できることが必要となる。

本稿は概念検索という同一概念の文書を検索する手法を提案する。具体的には、ユーザが検索システムにキーワードを入力したとき、そのキーワードを含まなくても、そのキーワードの背後にある概念と同じ概念を持つ文書が検索できるようになる。この手法は見かけ上、キーワード検索であるが、ユーザの入力したキーワードが背後に持つ概念と同じ文書が検索できるようになる。提案手法は文書を言葉のゆらぎや多義語などのノイズを含んだ単語集合、その文書の背後にある概念を情報の発生源と見なす。そこでノイズのあるデータからそのデータの発生源を特定するという問題に帰着できるので情報論的アプローチを提案手法に取り入れる。文書から概念を推測することによって、文書を単語空間でなく概念空間に配置することができる。概念空間における文書の位置関係は単語空間よりも文書の本質的な内容の類似性を反映している。その結果、本稿は提案したモデルを適用することによって、文書を概念に基づいて検索を行えることを実験によって証明した。

本研究は、検索の効率化のために、情報論的アプローチを利用することで文書を概念の類似性に基づいて検索する手法を提案する。論文の構成は次のようになっている。2 章では既存研究を振り返り、3 章では手法を実現する為に文書から利用する情報と、概念を抽出するモデルを提案する。4 章では実験を行い、5 章でまとめる。

2 既存研究

文書分類には分類する文書の種類によって大きく二通りに分けられる。教師付き文書の

分類と教師なし文書の分類がある。教師はテキスト文書の分類されるカテゴリである。本研究は後者の教師なしデータの分類を対象とする。この分野において代表的な手法に K-Means などのクラスタリング手法がある。これらの手法は文書をベクトル空間における類似関係を利用している。ベクトル空間モデル(Vector-Space Model:VSM)[1]は、単語を軸とする空間において、各文書を配置するモデルである。ベクトル空間における文書の座標は、軸に対してそれが含む単語の重みを座標とする。分類する文書集合に含まれる単語の総数を n とすると、空間の軸の数は n となる。

文書の分類を、出現した全ての単語でなくより少数の情報で分類するアプローチがある。その代表的なものが LSA (Latent Semantic Analysis)[2] である。LSA は特異値分解(singular value decompostion;SVD)に基づいた手法である。この手法の主張は LSA により文書集合は元々の軸よりも少数の軸で構成される空間で文書間の類似性が測定できるということである。ベクトル空間モデルにおいて空間は単語を軸とした空間で構成されるのに対し、LSA での空間は潜在的な意味空間である。意味空間を構成する軸は、ベクトル空間の軸を合成した軸である。この次元の縮小において SVD が用いられる。SVD を用いることで、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。

LSA は SVD によって導か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。一方、概念検索は文書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分類・検索が期待できる。

3 提案手法

本稿は文書に出現した単語でなく、文書の背後にある概念の類似性を推定し、その類似性によって分類を行う。多くの情報検索において前章で挙げたベクトル空間モデルが利用されている。ベクトル空間において文書が配置され、空間における類似度により文書の類似度が求められる。ベクトル空間モデル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

提案手法もベクトル空間モデルに基礎を置いているが、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概念で構成されている。従って、文書が単語を軸とするベクトル空間から概念を軸とする概念空間に配置される。その結果、概念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。文書の集合から概念の軸を推定するために、以下で文書・単語行列とモデルの定式化を行う。

3.1 文書・単語行列

単語の属性値を決定すれば、文書集合は文書・単語行列の形式で表現できる[3]。文書・単語行列 A は文書を行、単語を列とする行列である。文書を d_i、単語を w_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。文書 d_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。

3.1.1 局所的重み付け(Local Weighting)

局所的重み付けは文書 d_i 内の単語に対してのみ重み付けを行う

L1:出現頻度

$$P_{ij} = \frac{C(w_{ij})}{C(w_j)}$$

P_{ij}:文書 d_i における単語 w_{ij} の出現頻度
$C(w_j)$:文書 d_i に出現した単語の総数
$C(w_{ij})$:文書 d_i に出現した単語 w_j の数

L2:正規化されたエントロピー

$$H_{ij} = -\frac{1}{\log M}\sum_{j=1}^{M} p_{ij}\log p_{ij}$$

M:文書 d_i に出現した単語の総数
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。

$$\mathbf{G_{ij}=1-H_{ij}}$$

3.1.2 大局的重み付け(Global Weighting)

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

$$P_j = \frac{C(w_j)}{C(w)}$$

P_j:文書全体における単語 w_j の出現頻度。
$C(w_j)$:文書集合に出現した単語 w_j の数
$C(w)$: 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

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

N:文書集合に含まれる文書の総数
P_{ij}:文書 d_i 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。

$$\mathbf{G_j=1-H_j}$$

G3：文書数の逆数

$$H_j = -\log\frac{C(d)}{C(d_j)}$$

C(d):文書集合に含まれる文書数
$C(d_j)$:単語 d_j を含む文書の数

3.1.3 文書間の類似度

文書間の類似度はベクトル空間での類似度である。単語の重みをする文書ベクトルの cosine 関数を用いる。文書 d_i と文書 $d_{i'}$ の類似度は次のように求められる。

$$\cos ine(d_i, d_{i'}) = \frac{\sum_M a_{ij}a_{i'j}}{\sqrt{\sum_M (a_{ij})^2}\sqrt{\sum_M (a_{i'j})^2}}$$

この場合の類似度は単語を軸とするベクトル空間における文書の類似度である。ベクトル空間を構成する軸や、文書ベクトルにおける単語の重みが変化すれば、同じ文書間でもこの類似度も変化する。

3.2 提案するモデル

文書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文書は

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単語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。

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

a_{ij}:文書 d_i における単語 w_j の観測値
先に挙げた文書・単語行列 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。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
$m \leqq j$:概念の個数は単語の総数よりも小さい
以上より、文書・単語行列 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。

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 d_j 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) [4],[5]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。

$$SC(A|m) \cong -\log P_a(A)+\frac{m}{2}\log n$$

$P_a(A)$:文書・単語行列 A の最尤推定量
m:概念の個数
n:文書・単語行列 A の要素数

最尤推定量 $P_a(A)$は次の計算で求める。

$$P_a(A)=P_m(W)P_w(C)$$

$P_m(W)$:概念の個数が m の時の因子負荷量の最尤推定値
$P_w(C)$:因子付加量 W によって求められる因子得点の最尤推定値

最尤推定量 $P_a(A)$の値は EM アルゴリズム[6]で求める。EM アルゴリズムは観測されたデータから内部の状態を一意的に決定できない為に、直接、最尤推定を用いることができない場合に有効な推定方法である。

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。

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

rank(A)=r であるとき、U は N×r 行列、V は M×r 行列で、それぞれ $U^TU=I, V^TV=I$(I は単位行列)を満たす。Σ は A の特異値を含む r×r の対角行列である。LSA において行列 A は Σ において特異値の大きい順に k 個を選んで再構成された場合、その行列 A_k をとすると次のような形に書き直せる。

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

この再構成において文書・単語行列は(N×M)型行列から(N×m)型行列へと次元の縮小が行

われたことが分かる。この A_k は rank(A)=k の中で最もよい近似になることが知られている。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、少なければ元の文書・単語行列 A との誤差が大きくなる。

後者のモデル選択の指標に関しては SC を適用することで解決が出来る。

4 評価実験

4.1 実験の目的

提案した手法の概念検索の実現性について、次の二点で評価する。一つは概念の推測の可能性である。研究の目的の一つは文書集合に出現した単語からその概念を推測することである。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。次に抽出された概念を軸とする概念空間における文書の類似度を評価する。両者とも評価基準として再現率と適合率を用いる。再現率は文書集合中において該当する文書のうち、検索される文書の割合、適合率は検索された文書のうち、該当する文書の割合である。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 [7] 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。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。

4.2 概念の抽出性

提案したモデルにより文書・単語行列の重み付けの異なる文書・単語行列から抽出された軸を比較評価する。軸の評価は次の手順で行う。各々の軸における因子負荷量が高い順に単語を選ぶ。これらの単語は三つ以上の文書に出現しているので、それらを用いて文書の検索をすることが可能である。その検索によってし各カテゴリにおける再現率、適合率を平均したものを表 1 に示す。

重みは局所的重みと大局的重みを組み合わせて用いることで、軸と概念の対応関係が明らかになった。特定の概念については再現率と適合率を両方あげることができたが、他の軸では減少した。一つの軸が一つの概念に対応することが確認できた。表 1 は統計学の概念に相当している。今回の実験では L2* G3 によるものが再現率、適合率ともに高かった。従来、情報検索の分野で用いられることが多い tf·idf はここでは L1* G3 に相当する。統計学のカテゴリにおいて再現率、適合率は高い精度を示すが、他のカテゴリでも逆に高い精度を示した。これは一つの軸が特定のカテゴリに対応するという本研究の目的とは異なる。次にそれぞれの軸で因子負荷量が高かった順に単語を 10 示す。

表 1:重み毎による推定された軸の評価

重み	情報理論		情報検索		統計学	
	再現率	適合率	再現率	適合率	再現率	適合率
L1	0.5	10.0	0.5	12.6	42.0	92.8
L2	0.5	14.3	0	0	29.5	93.2
G1	3.8	58.5	5.5	7.8	41.0	86.3
G2	1.0	1.9	0	0	46.8	96.1
G3	0.5	0.8	0	0	46.8	96.3
L1*G1	2.5	4.7	3.5	6.3	36.5	89.0
L1*G2	0.5	1.0	0.5	12.6	42.5	97.8
L1*G3	0.5	1.0	0.5	12.6	42.5	97.8
L2*G1	4.0	5.8	5.5	7.8	41.0	86.3
L2*G2	1.0	1.7	0	0	47.0	98.3
L2*G3	0.5	0.8	0	0	47.0	99.2

単位は%

表 2:同一概念から出現した単語

重み付け	L1*G1	L1*G2	L1*G3	L2*G1	L2*G2	L2*G3
単語	単純	単純	単純	単純	単純	単純
	特殊	任意	任意	これら	標本	標本
	これら	下記	下記	特殊	項	項
	誤差	σ	σ	適切	変数	変数
	適当	項	項	様々	推定	推定
	条件	母	母	多次元	共	共
	多次元	他方	他方	条件	母	一定
	m	着目	着目	誤差	一定	母
	最小	未知	期待	行列	別	誤差
	モデル	期待	未知	モデル	誤差	仮定

表 1、表 2 から抽出された軸が概念として“統計学”という概念を推測するのに適していることが確認された。

4.3 文書間の類似度

提案したモデルによって推測された概念の軸で構成される空間と、次元縮小を行わない元のベクトル空間、SVD によって導出された潜在的な意味空間における文書の類似性の比較を行う。それぞれの空間において cosine 関数を用いて検索を行い、検索結果における文書の再現率、適合率を評価した。検索結果は cosine 関数が 0.7 以上の文書を利用した。

大部分の検索システムで用いられる単語のベクトル空間における文書の検索は、適合率は非常に高い値を示すが、再現率は非常に低くなっている。他の二つの空間に比較して次元の数が、出現した全ての単語の異なり数 4692 と大きいために、言葉のゆらぎや多義語などの単語のノイズに弱くこのような結果になったと考えられる。単語の重みについては SVD と提案手法のどちらも大きな相違は見られなかった。だが、手法そのものの比較で見た場

合、提案手法の方が、再現率、適合率でやや上回っている。これは概念空間における文書の類似関係が SVD による空間よりも文書の内容の類似性を反映していることが原因と考えられる。

表 3:各空間における情報の検索結果

空間の種類	情報理論		情報検索		統計学	
	再現率	適合率	再現率	適合率	再現率	適合率
元の空間: L2G2	15.1	90.0	18.4	90.0	16.2	95.0
元の空間: L2G3	15.1	90.0	18.4	90.0	16.2	95.0
SVD: L2G2	65.7	80.5	34.6	62.0	45.0	80.5
SVD: L2G3	64.9	81.7	36.1	66.3	48.6	80.3
概念空間: L2G2	68.3	86.2	40.5	69.1	53.8	86.6
概念空間: L2G3	68.5	84.5	45.0	68.2	54.5	86.5

単位は%

5 考察・まとめ

本稿は文書の検索を、出現した単語でなく、文書の概念に基づいて行う手法を提案した。提案手法は情報論的アプローチにより言葉のゆらぎや多義語などのノイズを含んだ文書から文書の背後にある概念を推測する。この手法によって文書は単語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、従来の手法よりも文書検索の精度が向上したことが確認できた。その理由は概念空間の文書の類似関係は単語を軸としたベクトル空間や SVD による空間よりも文書の本質的な内容の類似性を反映していると考えられる。更に概念の抽出、検索の精度を上げるには、モデルについての仮定を変える必要がある。

計算の簡易性の為に、概念の空間の軸は独立である仮定を置いたが、今回の実験で用いた文書は内容には相関があると考えられる。それにも関わらず今回の実験で、概念の軸が明確に抽出され、検索の精度が良かったのは、次の理由が考えられる。一つは実験データとなる文書を揃えすぎたことで、単語のノイズの問題があまり無かったことであり、二つ目は利用する単語を絞り込んだ結果、文書・単語行列がそれほど大きくならずゼロ・スパースネスの問題があまり無かったためと考えられる。

参考文献

[1] Salton, G., McGill, M. J.: Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[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]北研二: 確率的言語モデル,東京大学出版会,1999.

[4] J. Rissanen.: Fisher information and stochastic complexity. IEEE Transactions on Information Theory, 42(1):40–47, January 1996.

[5]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.

[6]Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. B 39 1977.

[7]茶筅http://chasen.aistnara.ac.jp/index.html.ja