概念空間における文書分類

6W－2

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

1　はじめに

現在の検索システムの多くはキーワード検索を採用している。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、自然言語を用いて検索を行い、検索結果はその単語を含む文章のリストである。ユーザが検索ニーズを自然言語で表現するには困難を伴い、検索結果は必ずしもユーザのニーズを満たさない[1]。

本稿は内容が類似した文書を検索する手法を提案する。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として扱い、その文書の背後にある概念を推測する。概念は内容と同義でユーザ側から見ればニーズに相当する。推測された概念を軸とする空間で文書あるいは単語の近接関係は他の手法で求められる空間より概念の類似性を反映している。提案手法により、単語そのものは含まなくてもユーザのニーズに合致した文書検索が行えることを実験によって証明する。

2　空間における文書配置

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、文書集合を配置するモデルである。LSA(Latent Semantic Analysis)[2] は特異値分解(singular value decompostion;SVD)に基づき、出現した全ての単語より少数の因子で元の VSM における文書配置を近似的に表現する空間を求める。従来の LSA に関しての研究は文書の再現率に寄与するが、次の二点が欠けている。

(1)文書の概念と因子の相関
(2)因子数決定指標の不在

3　概念空間の抽出

本稿は文書に出現した単語からその背後にある概念を推測し、その概念を軸とする空間に文書を配置する。この空間を概念空間と呼ぶ。提案手法もベクトル空間モデルに基礎を置いているが、相違点は空間の軸が概念で構成されていることである。前章で述べた問題点を(1)に関しては因子分析、(2)に関しては確率的コンプレキシティ（Stochastic Complexity:SC）[3],[4]を用いて解決する。

3.1　文書・単語行列

単語の属性値を決定することで文書集合は文書・単語行列の形式で表現できる[5]。本稿は局所的重み付けとして次の正規化されたエントロピーを用いる。

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

M:文書 d_i に出現した単語の種類
P_{ij}:文書 d_i における単語 w_j の相対頻度

次の変形を行い、重み付けとして利用する。

$$\mathbf{L2=1-H_i}$$

大局的重み付けとして単語毎のエントロピーを用いる。

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

N:単語 w_j が含まれる文書数
P_{ij}:文書集合における文書 d_i に含まれる単語 w_j の相対頻度

$G2=1-H_j$ と変形して重み付けとして利用する。行列 A の(i,j)成分の重みはこれらを用いて $a_{ij}=L2*G2(0\leqq a_{ij}\leqq 1)$とし、属性値として利用する。

3.2　概念の推測

文書の概念を推測する為に「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」を仮定したモデルを設定する。LSA が文書を原因、因子を結果としたのに対し、本稿は文書を結果、原因としてその概念をモデルにより推測する。因子分析を用いた次のモデルを用いて概念を推測する。

$$\mathbf{A=WC+UV}$$

A:単語･文書行列
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
V:独自因子得点行列、（i×j）型行列
m:推定する概念の個数

因子間は独立である仮定を置き、因子負荷量の推定には主因子法を用いる。

因子負荷量の推定には独自部分の評価と概念の個数を SC により同時に決定する。本稿で概念の個数 m の決定に用いた SC は次のように式になる。主因子法によって求めた A の重み $\hat{a}_{ij}$ を出現確率として利用する。

$$SC(A|m) \cong -\sum_i \sum_j \hat{a}_{ij} + \frac{m}{2}\log n$$

n:行列の要素数(=i×j)

4 概念空間における文書の類似度

4.1 実験の目的

提案した手法の概念による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、次は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。今回の実験で用いた単語の品詞は名詞と未知語である。

4.2 概念の抽出性

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 10 個選ぶ。その単語を用いた検索した時の同一因子軸における各カテゴリ毎の再現率、適合率を平均したものを表 1 に示す。SC により最適な概念の軸の数は 3 となった。

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	
	R	P	R	P	R	P
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% P:再現率 R：適合率

L1:文書 d_i における単語 w_j の出現頻度
G1:文書全体における単語 w_j の出現頻度
G3：文書数の逆数

$$H_j = 1 + \log \frac{C(d)}{C(d_j)}$$

C(d):全文書数
$C(w_j)$:単語 w_j を含む文書の数

提案手法によって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、他の概念では減少した。各軸が一つの概念に対応することが確認できた。重みとしては L2*G2, L2*G3 がほぼ同様の有効性を示した。

4.3 文書間の類似度

空間毎の文書の類似性を分類結果内の類似性を用いて比較する。LSA、概念空間においてそれぞれ 8 個(=2{因子、概念の正負}＾3{因子の数})に分類する。各グループに含まれる文書のカテゴリでもっとも多いのものをそのカテゴリの文書集合とし、表 2 に結果を示す。概念空間が各カテゴリにおいて内容の類似した文書を分類できることが確認できた。

表 2: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	
	R	P	R	P	R	P
LSA	−	−	−	−	24.5	67.3
グループ数	0		0		8	
概念空間	17.5	76.3	15.2	65.6	27.8	76.4
グループ数	1		4		3	

5 まとめ

提案手法は概念の存在を仮定し、モデル選択として SC を用いて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した。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だけでなく内容の類似した検索が可能になっ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 LSA 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。

参考文献

[1]川前,青木,安田.:情報理論的モデルを用いた情報検索,信学技報 Vol.101,No.192, 2001.

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

[4]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.

[5]北研二.:確率的言語モデル,東京大学出版会,1999.