2つの類似尺度を利用する類似文書決定手法の検討

1P－5

松本 一則　青木 圭子　帆足 啓一郎　橋本 和夫
KDD研究所

1. はじめに

類似文書を検索するための検索モデルとしては，tf・idfに代表されるベクトル空間モデルと，検索要求に対する適合確率と非適合確率の比で検索要求文書との類似性を表す確率型モデルが有名である．確率型モデルの方が，類似度（距離）の値に対する意味が明確であること，岩山等[1]が示したようにクラスタリング時の精度が良さそうなことから，筆者らは確率モデルに重点をおいて研究[2]を進めている．

しかし，先の岩山等の類似尺度を用いた検索実験の結果を詳細に分析すると，類似度がかなり高いものはほぼ正解であるが，少し類似度が下がったものの中に正解でないものが多く，全体の検索精度を下げていることが分った．このため，単一の類似尺度でだけでは十分な検索精度を得ることが難しいと判断し，岩山等の尺度と多項分布に基づいた類似尺度を組み合わせる手法の検討を進めた．今回，2つの類似尺度を特徴量とする空間における正解・非正解の分布を調査した結果，検索性能の向上できる見通しを得たので報告する．

2. 使用した類似尺度の計算方法

本稿で使用する類似度の1つは岩山等の類似尺度であり，もう1つは多項分布モデルを用いたものである．両者は，文書 d_x と文書集合 c が与えられた時の類似度 $P(c|d_x)$ の計算方法が異なっている．

2.1 岩山の計算方法

岩山等の場合，$P(c|d_x)$ を以下のように計算する．

$$P(c|d_x) = P(c)\sum_t \frac{P(t|c)P(t|d_x)}{P(t)}$$

- $P(t|d_x)$: d_x での単語 t の出現確率．
- $P(t|c)$: c での単語 t の出現確率．
- $P(t)$: 全検索対象文書中での単語 t の出現確率．
- $P(c)$: d_x が c 中の文書が含まれる確率（以下では，1となる）．

2.2 多項分布の計算方法

多項分布の場合，$P(c|d_x)$ は以下のように計算する．

$$\begin{aligned} P(c|d_x) &= {}_N C_{n(t_1)} \times {}_{N-n(t_1)} C_{n(t_2)} \times \cdots \\ &\quad \times \{P(t_1|c)\}^{n(t_1)} \times \{P(t_2|c)\}^{n(t_2)} \times \cdots \\ &= \frac{N_x!}{n_x(t_1)! \times n_x(t_2)! \times \cdots} \\ &\quad \times \{P(t_1|c)\}^{n(t_1)} \times \cdots \end{aligned}$$

- N_x: d_x 中の全単語の出現回数の和．
- $n_x(t_i)$: d_x での単語 t_i の出現回数．

なお，この定式化を用いた文書検索の手法は見当たらない．

2.3 2文書の類似度の計算方法

岩山の方法にせよ多項分布の方法にせよ，$P(c|d_x)$ を求め方によらず，文書 d_x と文書 d_y の類似度 $Sim(d_x, d_y)$ は以下ようにして計算する．これは，d_x と d_y をマージしたクラスタ（文書集合）が検索された時の文書 d_x と文書 d_y が正解である事後確率となっている．

$$Sim(d_x, d_y) = \frac{P(\{d_x, d_y\}|d_x) \cdot P(\{d_x, d_y\}|d_y)}{P(\{d_x\}|d_x) \cdot P(\{d_y\}|d_y)}$$

3. 単一類似度による正解と非正解の分布

ここでは，まず岩山の手法で類似度を計算した場合における，正解文書（適合文書）と非正解文書（非適合文書）の分布について述べる．

分布を求めるための実験[3]は，1993年～1999年の公開特許公報の中から抽出した1万件の特許を検索対象にして行なわれ，21回の検索要求を行なった．各検索要求に対する1万件の特許に対する網羅的な類似性の判定は専門家によって行なわれた．結果を図1に示す．

図1では，類似性が高い領域（類似度が-1.0以下）では，ほとんど正解文書しか存在せず，ほぼ完全に正解と非正解を分離できる．しかし，一部の正解文書の類似度が良くないため，非正解文書とうまく分離できていない．特許文書からくる性質なのか不明ではあるが，正解文書の分布は非正解文書の分布に比べ，平坦で全体的に広がっていることがわかる．

"Documet Similarity Test Based on Two Criteria", Kazunori MATSUMOTO, Keiko AOKI, Keiichirou HOASHI and Kazuo HASHIMOTO: KDD R&D Laboratories Inc..

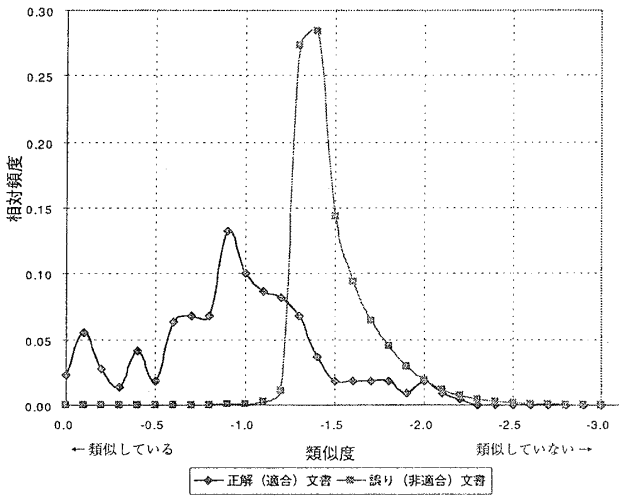

図 1: 正解文書と非正解文書の分布

2.2 の多項分布の類似尺度を用いて，同様のグラフを得たが，正解分布と非正解分布の分離性は図 1 より悪い状況であった．

4. 両類似度を特徴量にする場合

各類似度を使った場合の正解と非正解の分布状況が分ったので，各類似度を特徴量とする 2 次元の特徴空間における，正解文書と非正解文書の散布図を各検索要求毎に得た．すると，全ての検索要求において，正解と非正解の文書が同様の分布のしかたを示した．

図 2 と 3 は，それぞれ，とある検索要求における散布図の例である．これらの図から，2 つの特徴量を合せて使用した場合，正解と非正解の分離が良くなって，検索精度を上げることが出来そうであることが分かる．

5. おわりに

従来から知られている岩屋の類似度以外に，新たに導入した多項分布の類似度を使用し，各類似度を特徴量とする 2 次元特徴空間上で正解と非正解の文書が高精度に分類できる見込を得た．

最後に、日頃御指導頂くＫＤＤ研究所村谷所長、知識情報処理グループの各位に感謝します．

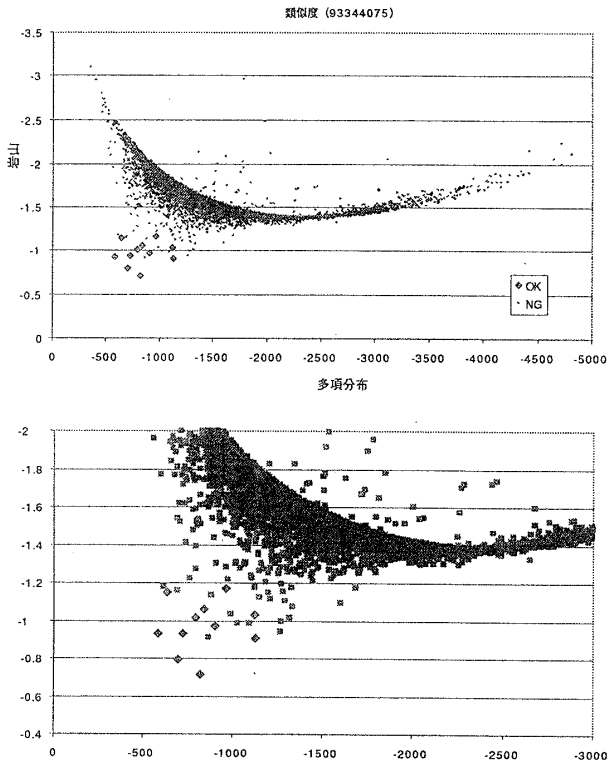

図 2: 2 つの類似度を用いる場合の例 (1)

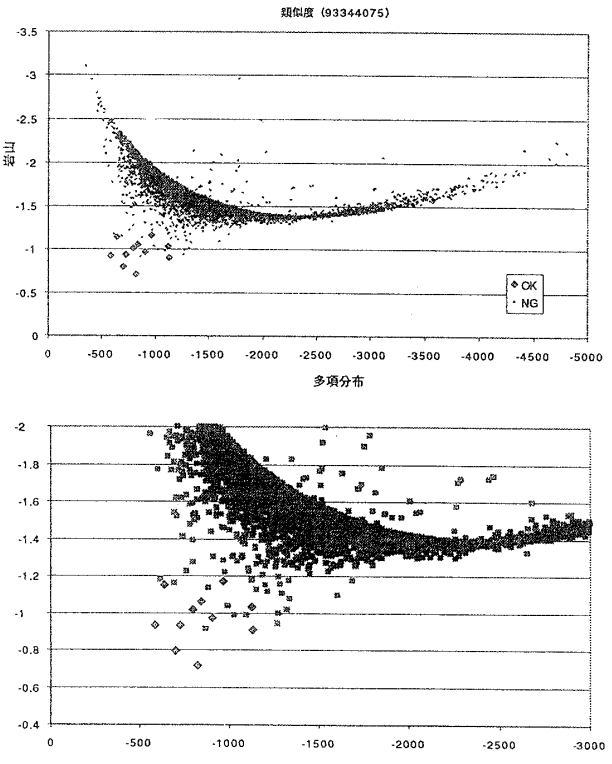

図 3: 2 つの類似度を用いる場合の例 (2)

参考文献

[1] Makoto IWAYAMA, Takenobu TOKUNAGA, “Hierarchical Bayesian Clustering for Automatic Text Classification”, Proceedings of IJCAI-95, pp.1322-1327, 1995.

[2] 青木, 松本, 橋本, “類似ドキュメントの発見手法の検討”, 第 54 情処全大, 3-39, 1997.

[3] 青木, 松本, 橋本, “類似検索を応用した特許通知システムの試作”, 第 58 回情処全大,2U-2, 3-145, 1999