社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－MPS－44 (5)
2003／5／9

類似テキスト検索のための多重トピックテキストモデル

上田 修功　斉藤 和巳
日本電信電話株式会社　NTTコミュニケーション科学基礎研究所
〒619-0237　京都府相楽郡精華町光台2-4
{ueda, saito}@cslab.kecl.ntt.co.jp

本稿では，あるテキストに対し概念的に類似したテキストを検索する手法を提案する．提案法は，テキストを単語の集合として捕えるという点で従来法と共通するが，テキストの確率モデルを土台とするという点で従来法と本質的に異なる．提案法では，各文書を確率モデルに基づいて推定したトピックベクトル空間上で文書間類似度を算出する．トピックベクトルの推定アルゴリズムは，高速かつ解の大域的最適性を理論保証する．検索結果に対する妥当な定量的評価基準を新たに導入し，webページを用いた検索実験を通して提案法の従来法に対する顕著な優位性を示す．

Multi-topic Text Model for Topic-based Text Retreival

Naonori Ueda, Kazumi Saito
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{ueda, saito}@cslab.kecl.ntt.co.jp

This report proposes a novel method for text retrieval based on probabilistic multi-topic text model. In the proposed method, document similarity measure is computed by using latent topic vectors which can be estimated by using our probabilistic model. The derived estimation algorithm is quite efficient and theoretically guaranttees the global optimality of the estimate. Through experiments using real web pages, we show that the proposed method could significantly outperform the conventional methods.

1 はじめに

近年，webページをはじめ膨大なオンライン文書が蓄積されつつある．文書の自動分類技術や検索技術はこうした大量文書を知識源として有効利用するための重要要素技術と位置づけられる．筆者らは，先に，テキストを多重のトピックに自動分類するための確率モデル(パラメトリック混合モデル:PMM)を考案し，webページの自動分類実験でその有効性を確認した[1].

テキスト分類は文書群を予め定めてトピッククラスに分類するタスクであり，テキスト検索で要求される同一クラス内での文書の類似性やクラス間での文書の類似性を直接評価できない．その意味でテキスト分類はテキスト検索と異なるタスクと言える．本稿では，PMMをテキスト検索モデルとして拡張し，テキスト検索における確率モデルアプローチの有効性を検証する．

2 従来法

テキスト検索手法の核は文書d_m, d_n間の類似度の定義にあると言える．$\boldsymbol{x}_n=(x_{n,1},\ldots,x_{n,V})$を$d_n$中の単語頻度ベクトル($x_{n,i}$は想定語彙集合$\mathcal{V}=\{w_1,\ldots,w_V\}$中の単語$w_i$が$d_n$中に出現した頻度を表す)とすると，従来法では，$d_m, d_n$間の類似度として$\boldsymbol{x}_n, \boldsymbol{x}_m$のコサイン類似度: $s_{m,n}=\cos(\boldsymbol{x}_m, \boldsymbol{x}_n)$が広く用いられる[2]．本稿ではCOS法と呼ぶ．更に，単語の重要度を考慮した重み付きコサンイン類似度も提案されている[3]．重みとして，単語の頻度の対数の逆数で定義されるIDF(Inverse Document Frequency)が用いられる．即ち，$\boldsymbol{x}$の第i要素x_iに対して$\log(s/s_i)$を乗じて重み付けする．$s(s_i)$は対象文書総数(w_iが出現した文書総数)を表す．これによりどの文書にも多く出現した単語の重要度が下げられる．本稿ではIDF法と