初期検索文書のトピック分布に基づく関連性フィードバックの一考察

芹澤 翠 [a]　　小林 一郎 [b]

お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻

a) *serizawa.midori@is.ocha.ac.jp*　b) *koba@is.ocha.ac.jp*

概要　本研究では，初期検索で得られた文書における潜在情報を考慮した擬似関連性フィードバック手法の提案を行う．潜在情報の抽出には確率的文書生成モデルの一つである潜在的ディリクレ配分法（Latent Dirichlet allocation）を利用し，それによって得られた文書のトピック分布を検索対象文書モデルとして扱い，またこの分布に基づいたフィードバックを作成することで潜在情報を考慮した関連性フィードバックを試みた．実験では，文書検索課題を用いて提案手法と表層情報を用いた手法を比較し潜在情報を用いた手法の有効性について検証を行い，ユーザの要求情報が少ない場合には潜在情報が有効であることが分かった．

キーワード　潜在トピック，関連性フィードバック，Latent Dirichlet allocation

1　はじめに

情報検索においてユーザが欲しい情報を得るのは容易でない．これにはユーザが十分な知識がなく的確なクエリ作成ができない等の理由が考えられる．これを克服する手段として関連性フィードバック (Relevance feedback:RF) の研究が広くなされている．RF とは入力されたクエリにより収集された初期検索結果の文書が要求に関連しているか否かの情報を用いて元のクエリを更新し，より良い検索結果を得るための手法であり，関連判定を人手で行う明示的 RF と初期検索結果上位の文書を関連文書として判定する擬似 RF とがある．手法としては，ベクトル空間モデルをベースとする手法や言語モデルをベースとする手法 [1] などがある．本稿では，潜在的ディリクレ配分法 (Latent Dirichlet allocation:LDA)[2] を利用し文書内の潜在トピックを考慮した擬似 RF 手法を提案し，実験において表層情報のみを用いた手法との比較を行い，潜在情報の RF における有効性について考察する．

2　潜在情報を考慮した関連性フィードバック

LDA は確率的文書生成モデルの一つであり，文書 d はトピックの多項分布 θ_d として，トピック j は単語の多項分布 ϕ_j として表現される．また文書 d での単語 w の出現確率は，$\theta_{d,j}$ を文書 d でのトピック j の出現確率，$\phi_{j,w}$ をトピック j での単語 w の出現確率，T をトピック数として，$\sum_{j=1}^{T} \theta_{d,j}\phi_{j,w}$ $\cdots(*)$ と表現できる．

本研究では，LDA により推定された文書のトピック分布を用い，初期検索文書のトピックを考慮してクエリを更新し文書を再ランキングする．具体的な手続きは以下の通りである．

step 1. 初期検索

単語出現確率の最尤推定値にスムージングを施したユニグラム言語モデル [3] を用いて表現されたクエリと文書を対象に KL ダイバージェンス (KLD) 検索モデル [1] により検索された初期検索結果上位 m 件を再ランキング対象文書 D とする．

step 2. トピックベースモデルの構築

LDA によりクエリと文書群 D のトピック分布を推定し，クエリのトピック分布をクエリモデル θ_q とし，D についても同様にトピックベースのモデルへ変換する．また，D の内上位 n 件を関連文書と見なし，関連文書の平均トピック分布をフィードバックモデル θ_F とする．

step 3. クエリ更新と文書再ランキング

パラメータ $a(0 \leq a \leq 1)$ を導入して，以下の式により新しいクエリモデル $\theta_{q'}$ を作成する．

$$\theta_{q'} = (1-a)\theta_q + a\theta_F \qquad (1)$$

このクエリを用いて KLD により D を再ランキングしたものを最終的な検索結果とする．

3　実験

実験では NTCIR-2 の情報検索システム評価用テストコレクションを用いた．評価は日本語検索課題 30 件を対象とし，各課題に対して約 1,400 文書を検索対象とした．クエリには検索課題の検索要求文 <DESCRIPTION> を用い，再ランキング対象文書数 $m = 100$，関連文書と見なす文書数 $n = 10$ とした．初期検索時のスムージングパラメータ値は 1,000 とし，LDA の文書トピック分布とトピック-単語分布それぞれに対する事前分布のパラメータは $\alpha = 0.1$，$\beta = 0.1$ とした．

実験では (a) クエリ更新式 (1) の調整パラメータ a，(b) フィードバック作成に用いる文書 n 件での適合文書

数の割合（初期検索精度）をそれぞれ変更させて評価を行った．尚，(b) は，初期検索精度によるフィードバックの性能を調査するため行った．評価尺度には，P@10(ランキング上位 10 文書の適合率) と MAP(平均適合率の平均) を用いた．

3.1 実験結果

図 1 に提案手法の P@10 評価結果を示す．MAP の結果は余白の都合上省略する．加えて比較のため，2 章の step 2 において，トピックベースでなく初期検索と同様に言語モデルでクエリと文書を表現して検索を行った結果を図 2 に示す．このモデルは潜在情報は考慮していなく単語の出現頻度にのみ基づいて計算されるため，表層情報のみを用いた手法と考えられる．また実験では，トピック分布を用いた手法の他に，LDA により推定された単語出現確率 (∗) により文書を表現した手法でも実験を行ったが，トピック分布を用いた結果（図 1）と大きな差異は見られなかった．図 1,2 の各線は初期検索精度毎の P@10 の値を示している．

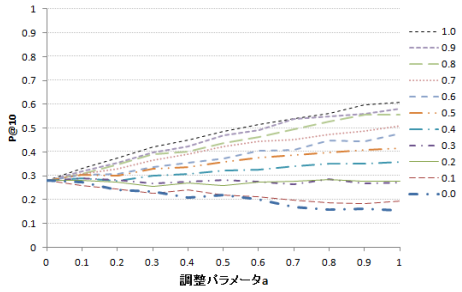

図 1 トピックベース手法の結果

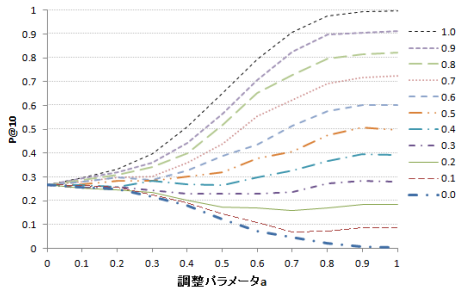

図 2 表層情報ベースの手法の結果

3.2 考察

結果を見ると，初期検索精度が低い場合と高い場合で傾向が異なることが分かる．初期検索精度が 0.0〜0.3 と低い場合は，表層ベース手法は調整パラメータ a が大きくなるにつれて精度は低くなり，フィードバックのみを考慮した場合 ($a = 1$) は初期検索精度を下回っている．一方，トピックベース手法はフィードバックの割合が多くなるにつれて精度は低くなっているものの，フィードバックのみを考慮した場合でも初期検索精度を上回っている．擬似 RF では，精度の低い検索結果から作成されたフィードバックはユーザの要求に関連する語は反映されにくいと考えられる．表層ベース手法では，初期精度が低いとユーザの要求を満たす単語への重み付けが低くなるため, 新しいクエリでの検索は初期精度を下回り，トピックを考慮した手法では，フィードバックにユーザの要求を直接表現する単語がなかったとしてもトピックを介してその要求に近い概念を持つ文書が検索され，初期精度を上回ったと推測できる．次に初期検索精度が 0.4 以上と高い場合，トピックベース手法はフィードバックの割合が多くなるにつれ精度は高くなり初期精度を上回っているが，表層ベース手法ほどの改善は見せていない．これはユーザの要求に直接関連しない語にも影響を受けてトピックが推定されるため，表層ベース手法に比べ要求の表現が曖昧であることが原因であると考えられる．

以上の考察から，表層ベース手法は直接的にフィードバックの精度に影響され，一方，トピックを用いた手法はフィードバックの精度に影響されにくいと分かった．これより，初期精度が良い場合は表層ベース手法に劣るが，初期精度が悪い場合トピックを考慮した手法は有効であると言える．また，RF は初期検索より良い結果をフィードバックの利用により得ることが目的であり，フィードバック反映後の精度が初期精度を上回っているのは，提案手法の初期精度 0.0〜0.2 のときのみであることから，潜在情報を用いた方法は初期結果を底上げする点で有効であることが分かる．

4 おわりに

本稿では潜在情報を考慮した RF 手法の提案をし，実験によりその有効性を表層情報を用いた手法と比較し検証した．その結果，ユーザの要求の情報が多い場合には表層情報を用いた手法の方が有効であるが，ユーザの要求情報が少ない場合にはトピックを考慮した手法が有効であることが分かった．今後の課題としては，フィードバック生成用の文書とユーザ要求との適合度の計り方や調整パラメータ等の各変数についての検証などが挙げられる．

参考文献

[1] Zhai, C. and Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval, In Proceedings of CIKM'01, pp.403-410, 2001.

[2] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, Journal of Machine Learning Research, 3, pp.993-1022, 2003.

[3] Zhai, C. and Lafferty, J.: A study of smoothing methods for language models applied to information retrieval, ACM Transactions on Information Systems, 22, 2, pp.179-214, 2004.