

潜在的意味を考慮した効果的な適合フィードバックへの取り組み

芹澤 翠 小林 一郎

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

{serizawa.midori, koba}@is.ocha.ac.jp

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ことが困難なことが多々ある．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．このことを克服する手段として適合フィードバック (Relevance feedback: RF) の研究が広くされている．RF とは入力されたクエリによる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たクエリを更新し、より良い検索結果を得る手法であり、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．

本稿では、潜在的ディリクレ配分法 (Latent Dirichlet allocation: LDA) [1] を利用し文書内の潜在トピックを考慮した擬似 RF 手法について表層情報を用いた手法との比較を通し考察を行う．また、この考察を踏まえ、文書内の局所的な情報を用いて生成されたフィードバックによる潜在情報考慮の適合フィードバック手法を提案し、実験によりその有効性について検証する．

2 関連研究

情報検索に潜在情報を考慮した手法は様々な提案されており、その有効性が確認されている．Wei ら [6] は情報検索を目的とした言語モデルの枠組みに則った LDA に基づく文書モデルを提案し、文書クラスに基づいたモデルよりも精度が良いことを示している．また、Yi ら [8] はトピックモデルの情報検索における有用性について調査し、LDA のようなモデルは文書検索に有効であることを示した．

RF 手法としては、Rocchio のアルゴリズム [5] に代表されるようなベクトル空間モデルを基礎とする手法や言語モデルを基礎とする手法 [9] などがあるが、RF に LDA を適用した手法も提案されている．Ye ら [7] は、クエリに関連するトピックを用いてトピックに基づくフィードバックモデルを提案している．Harashima ら [3] は、フィードバックと各文書を対象に LDA により推定されたトピックを用い、フィードバックと類似し

たトピック分布を持つ文書が検索上位に現れるよう工夫している．

また、クエリ拡張のためのフィードバック作成源に着目し、その構成要素を文書でなくより細かい単位とする研究もされている．これによりクエリ拡張の際に不要な語が追加されることを防ぐことができ、精度の向上が期待できる．具体的には要約 [4] や文 [2] などが用いられており、その有用性が示されている．

3 潜在情報を考慮した RF

3.1 潜在トピックを考慮した RF 手法

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、初期検索結果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．

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．

3.1.1 言語モデル

文書とクエリを表現する言語モデルとして、単語を素性とし、その重みを単語出現確率の最尤推定値に Dirichlet スムージング [10] を施した値とするモデルを用いる．単語 w_j の文書 t における最尤推定値と、その Dirichlet スムージングはそれぞれ次式 (1), (2) のように計算される．

$$P_t^{mle}(w_j) = \frac{tf(w_j, t)}{|t|} \quad (1)$$

$$P_t^{dir}(w_j) = \frac{tf(w_j, t) + \mu P_{D_{all}}^{mle}(w_j)}{|t| + \mu} \quad (2)$$

ここで、 D_{all} は対象文書群全体を表しており、 μ はスムージングパラメータである．

3.1.2 提案手法における処理の流れ

具体的な手続きは以下の通りである．

Step 1. 初期検索

3.1.1 節の言語モデルにより表現されたクエリ P_q^{mle} と文書 P_D^{dir} を対象に、式 (3) で計算される KL ダイバージェンス (KLD) 検索モデル [9] を

用いて検索された検索結果上位 m 件を再ランキング対象文書 D とする。

$$KL_score(\mathbf{d}, \mathbf{q}) = -KL(P_{\mathbf{q}}^{mle}(\cdot) || P_{\mathbf{d}}^{dir}(\cdot)) \quad (3)$$

$$KL(P_{\mathbf{q}}(\cdot) || P_{\mathbf{d}}(\cdot)) = - \sum_w P_{\mathbf{q}}(w) \log P_{\mathbf{d}}(w)$$

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 のトピックを推定し、クエリをトピックベースモデル $P_{\mathbf{q}}^{tpc}$ へ変換し、 D についても同様に変換する。トピックベースモデルでは (1) トピック分布 θ_d , (2) 推定された単語出現確率 $\sum_{j=1}^T \theta_{d,j} \phi_{j,w}$ を用い 2 通りに文書を表現する。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 $P_{\mathbf{F}}^{tpc}$ 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、以下の式により新しいクエリモデル $P_{\mathbf{q}'}^{tpc}$ を作成する。

$$P_{\mathbf{q}'}^{tpc} = (1-a)P_{\mathbf{q}}^{tpc} + aP_{\mathbf{F}}^{tpc} \quad (4)$$

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。

3.2 実験

3.2.1 実験仕様

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。また、初期検索時のスムージングパラメータ値 μ は 1,000 とし、LDA のトピック数は予備実験により $K = 50$ とし、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 $\alpha = 50/K$, $\beta = 0.01$ とした。

実験では (a) クエリ更新式 (4) 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、ランキング上位 10 文書の適合率 $P@10$ と平均適合率の平均である MAP を用いた。比較する手法は初期検索と次の 3 手法である。

- トピックベースの手法
 - 文書を文書固有のトピック分布で表現した手法 (TOPIC)
 - 文書を推定された単語出現確率で表現した手法 (TPCWORD)

- 表層ベースの手法

- 文書を初期検索と同様に言語モデルで表現した手法 (WORD)

3.2.2 結果と考察

TPCWORD, WORD の初期検索精度別に調整パラメータ a を変更させた $P@10$ 評価結果をそれぞれ図 1, 2 に示す。TOPIC についても実験を行ったが、TPCWORD の結果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。また、各手法の評価結果が最も良かった a の値でのスコアを表 1 に示す。

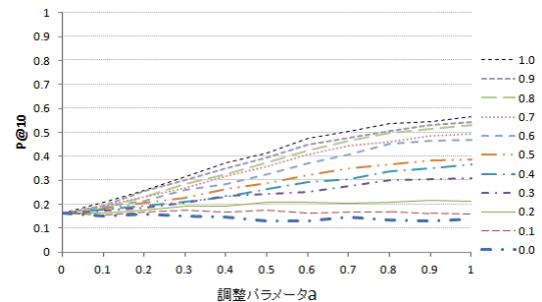


図 1: トピックベース手法の結果

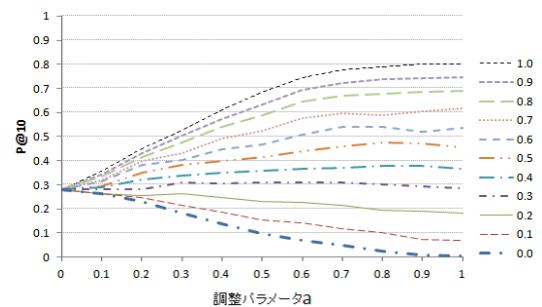


図 2: 表層ベースの手法の結果

表 1: 全課題を対象にした各手法の評価結果

	P@10	MAP
初期検索	0.331	0.239
TOPIC ($a = 0.9$)	0.283	0.225
TPCWORD ($a = 1.0$)	0.286	0.215
WORD ($a = 0.9$)	0.348	0.239

図 1, 2 から、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 a が大きくなるに従い精度は低くなり、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従い精度は低くなっているが、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。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っているが、表層ベース手法ほどの改善は見せていない。図 1, 2 から、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。評価についても初期検索精度

の低い課題¹のみを対象に見てみると、結果は表 2 の通りである。この結果から、表 1 の全課題に対する評価では、トピックを考慮した手法は表層情報のみを用いた手法より低い結果となっていたが、初期検索の精度が低い場合にはトピックを考慮した方が精度が良くなることが分かった。

表 2: 初期検索精度の低い課題の結果

	P@10	MAP
初期検索	0.094	0.112
TOPIC ($a = 0.2$)	0.141	0.118
TPCWORD($a = 0.0$)	0.124	0.106
WORD ($a = 0.0$)	0.124	0.116

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。初期検索精度が高い場合はユーザの要求に直接関連しない語から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になり精度が低くなったと考えられる。

4 局所情報を用いた潜在情報を考慮した RF

4.1 潜在情報を考慮した手法からの改善点

前章の考察から、検索精度を上げるための改善手法の一つとして、フィードバック生成源の単位を文書より細かいものとする事が挙げられる。これにより不要な情報を避けることができ推定されるトピックが限定されるため、精度が向上することが考えられる。今回は、フィードバック生成源を文とし、対象文書群から表層情報に基づき選択した文集合をフィードバックとして利用することとした。

4.2 局所情報を用いた手法

4.2.1 フィードバックの作成方法

対象文書群の各文書内の文を対象に、以下の式²を用いて文をランキングする。

¹今回は初期検索精度が 0.4 以下のものとした。該当する課題数は 18 課題である。

²予備実験により数手法比較した中で一番結果が良かったものを採用した。比較した他の手法は、(i) 語の頻度 [2] (ii)tf-idf を用いたコサイン類似度 (iii)BM25 (iv) t_q に重複を許した単語を用いたもの (v) 式 (5) の分子を二乗したもの [4] (vi)3.1.2 節の初期検索と同じ手法の 6 手法である。

$$score(s_i) = \frac{t_q(s_i)}{n_q} \quad (5)$$

ここで $t_q(s_i)$ は文 s_i 内に含まれるクエリ内語彙数、 n_q はクエリ内語彙総数である。この内、高いスコアを持つ文³を結合し一文書と見なしたものをフィードバックとする。

4.2.2 提案手法における処理の流れ

3.1.2 節の Step 2 を以下のように変更する。

Step 2. フィードバックとトピックベースモデル構築

Step 1 で得られた文書群 D を対象に、4.2.1 節に記載されたように文集合を取得する。これを一文書とみなし、クエリと文書群 D に加えてトピックを推定し、トピックベースモデルへ変換する。

5 実験

5.1 実験仕様

実験に用いるテストコレクション、課題、クエリ、パラメータはすべて 3.2 章の実験と同様にし、再ランキング対象文書数は $m = 100$ とした。評価尺度には、P@10 と MAP を用いる。比較する手法は、初期検索と 3.2.1 節での比較手法においてフィードバックに文を用いた、s-TOPIC、s-TPCWORD、s-WORD の 3 手法とする。

5.2 実験結果

全課題を対象にした各手法の評価結果が最も良かった a の値でのスコアを表 3 に示す。3.2 章で一番精度の良かった WORD も比較のため記載した。

表 3: 全課題を対象にした各手法の評価結果

	P@10	MAP
初期検索	0.331	0.239
s-TOPIC ($a = 0.8$)	0.372	0.253
s-TPCWORD($a = 1.0$)	0.369	0.255
s-WORD ($a = 0.9$)	0.293	0.245
WORD ($a = 0.9$)	0.348	0.239

また、初期検索精度の低い課題と高い課題⁴のトピックを考慮した手法の結果をそれぞれ表 4、5 に示す。比較のため、フィードバック生成源を文書とした TOPIC と TPCWORD の結果も記載する。

表 4: 初期検索精度の低い課題の結果

	P@10	MAP
初期検索	0.094	0.112
s-TOPIC ($a = 0.8$)	0.423	0.253
s-TPCWORD($a = 1.0$)	0.415	0.255
TOPIC ($a = 0.2$)	0.141	0.118
TPCWORD($a = 0.0$)	0.124	0.106

³今回は、予備実験により、順位が 1 位の文のスコアに 0.7 を掛けた値以上のスコアを持つ文とした。

⁴課題の選び方は 3.2.2 節と同様にした。高い課題は低い課題に該当しなかった課題とした。

表 5: 初期検索精度の高い課題の結果

	P@10	MAP
初期検索	0.667	0.419
s-TOPIC ($a = 0.9$)	0.527	0.382
s-TPCWORD($a = 1.0$)	0.518	0.386
TOPIC ($a = 0.9$)	0.642	0.407
TPCWORD($a = 1.0$)	0.650	0.432

5.3 考察

表 3 から, トピックを考慮した手法 s-TOPIC, s-TPCWORD は他手法より良い精度となったことが分かる. これは, フィードバック生成源を文にしたことにより要求に含まれる概念を持つ単語がフィードバックに多く含まれたため, 新しいクエリの持つトピックが明確になり, 関連文書との類似性が増したことによると考えられる. また, s-WORD と WORD を比べると s-WORD は MAP は上回っているが P@10 は WORD より低い結果となっている. s-WORD は文書を語の頻度に基づいて表現しているが, フィードバックを構成している文を語彙の豊富さで取ってきているため, フィードバック内ではクエリの語の頻度が極端に高くなってしまいう可能性がある. このことにより, フィードバックと関連文書の類似度が下がり, s-WORD の方が値が低くなったと推測できる. しかし, MAP を見ると, 文をフィードバック生成に用いた 3 手法は文書を用いた手法よりもすべて上回っており, 文を用いたフィードバックは有効であることが分かる.

次に, 初期検索精度別の結果を見ると, 表 4 から精度の低い課題はフィードバックに文書を用いた手法よりも大幅に精度が向上している. また, 文書を使った手法では a は 0.0 に近い値を取っている一方, 文を使った手法では 1.0 に近い値を取っている. a は値が大きいほどフィードバックを考慮しているということなので, この結果からも文から作成されたフィードバックが有効であることが分かる. 他方, 精度の高い課題については表 5 から精度が下がっていることが分かる. この原因は, まず, 初期検索精度が低い検索で良い結果が出れば初期検索精度の高い検索でも良い結果が出るという仮定の下, 文選択方法を決めるための予備実験において初期検索精度の低い課題を用いて決定した方法を使用したことが考えられる. また, 今回は文の選択方法にトピックの多様性などは考慮していないため, 類似した内容の文ばかりが選択されてしまったことも原因として考えられる. このことに関しては, さらなる調査と考察が必要である.

6 おわりに

本稿では潜在情報を考慮した適合フィードバック手法についての考察と, それを踏まえて, 文から生成さ

れたフィードバックを用いた潜在情報を考慮した RF 手法を提案し, 実験によりその有効性を検証した. その結果, フィードバック生成に文書を用いた手法や潜在情報を考慮しないで文をフィードバック生成に用いた手法よりも提案手法の精度が良く, 文から作成されたフィードバックと潜在情報を用いた適合フィードバックを組み合わせる手法が有効であることが分かった.

今後の課題としては, 今回は文選択にトピックの多様性は考慮していないため, クエリの持つ話題を考慮した文選択への取り組みや, 他のデータセットでの実験などを考えている.

参考文献

- [1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993-1022.
- [2] Debasis Ganguly, Johannes Leveling, and Gareth J. F. Jones. 2011. Query expansion for language modeling using sentence similarities. In *Proceedings of the Second international conference on Multidisciplinary information retrieval facility (IRFC'11)*, 62-77.
- [3] Jun Harashima and Sadao Kurohashi. 2011. Relevance Feedback using Latent Information, In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 1037-1045.
- [4] Adenike M. Lam-Adesina and Gareth J. F. Jones. 2001. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*, 1-9.
- [5] Joseph John Rocchio. 1971. Relevance feedback in information retrieval, In *The SMART Retrieval System: Experiments in Automatic Document Processing*, 313-323.
- [6] Xing Wei and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*, 178-185.
- [7] Zheng Ye, Jimmy Xiangji Huang, and Hongfei Lin. 2011. Finding a good query-related topic for boosting pseudo-relevance feedback. *J. Am. Soc. Inf. Sci. Technol.* 62, 4 (April 2011), 748-760.
- [8] Xing Yi and James Allan. 2008. Evaluating topic models for information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*, 1431-1432.
- [9] Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management (CIKM '01)*, 403-410.
- [10] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179-214.