潜在的な意味の関連性に基づく文書処理手法の開発

理学専攻・情報科学コース　芹澤 翠 (指導教員：小林 一郎)

1　はじめに

近年，文書など離散データの解析手法として，文書の生成過程を単語出現確率の変動を元にトピックを考慮しモデル化したトピックモデルが広く用いられている．トピックモデルの代表例としては，pLSA や LDA[1] などがあり，情報検索，文書要約や文書分類などに適用されている．トピックモデルの特徴は一文書に複数トピックが存在することを表現できることであり，各文書は固有のトピック比率 θ_d として，各トピックは固有の単語分布 ϕ_j として表現される．本研究では，文書の持つ潜在トピックの関連性に着目し，トピック追跡と適合フィードバック (Relevance feedback:RF) へのトピックモデルの適用を行う．前者ではトピック間類似度を用いたトピック抽出と追跡手法を，後者ではフィードバック作成と文書検索にトピックを考慮した手法をそれぞれ提案し，実験によりその有効性を検証する．

2　文書内トピック数を考慮したトピック追跡

2.1　概要

LDA を用いて抽出されたトピックを追跡し，時間変化に伴う話題の変化を解析することを目的とする．LDA を用いるには文書内トピック数を与える必要があるが，事前に決めることは困難である．そのため，トピックの内容の類似度によりトピック数を決定した上でトピック抽出を行い，その類似度を用いて追跡する．

2.2　手法

トピック数判定とトピック追跡では，トピックを単語を素性とするベクトルで表現し，その類似度によりトピック間の関連度を測る．トピック内の語の特徴量には，tf-idf 値の文書をトピックに置き換えた term-score を用いる．

$$\text{term-score}_{k,v} = \hat{\beta}_{k,v} \log \left(\frac{\hat{\beta}_{k,v}}{(\prod_{j=1}^{K} \hat{\beta}_{j,v})^{\frac{1}{K}}} \right) \qquad (1)$$

ここで，$\hat{\beta}_{k,v}$ は推定された語 v のトピック k での出現確率，K はトピック数を表している．また，トピック間類似度指標にはコサイン尺度を用いる．

2.2.1　トピック数の判定

トピック数に大きめの値を意図的に与えて抽出したトピックに対し，閾値以上の類似度を持つトピック組を同じ内容を持つ‘関連トピック’と見なす．その中に含まれていないトピックを‘単独トピック’とみなし，関連トピックを 1 つのトピックとしてまとめることで，‘結合トピック’を生成する．複数の結合トピックに含まれるようなトピック（‘重複トピック’と呼ぶ）を主張性の弱いトピックと捉え,「単独トピック数」と「重複トピックを除いた結合トピック数」の和をその文書に潜在するトピック数と判定する．

2.2.2　トピック追跡

対象期間の各日において，2.2.1 節に記載した方法により判定したトピック数を用いて LDA によりトピックを抽出する．各日のトピック集合を対象に，連続する 2 日間の各トピック間の類似度が閾値以上ならばトピック間に関連があるとし，これを対象期間分繰り返す．

2.3　実験

2.3.1　実験仕様

対象とする文書はニュースサイト「YOMIURI ONLINE（読売新聞）」,「毎日 jp（毎日新聞）」からキーワード「尖閣」を与えて収集した 2010 年 11 月 13 日から 15 日までの 86 件のニュース記事である．トピック数判定の際に最初に与えるトピック数は 18 とした．

2.3.2　結果と考察

トピック数判定の結果を表 1 に示す．比較の為，HDP-LDA により推定されたトピック数と推定されたモデルのパープレキシティが低くなったトピック数 (「LDA」と記載) も計算した．また，トピック抽出と追跡の結果の一部を表 2 と図 1 に示す．

表 1: 各手法により判定されたトピック数

日にち	提案手法	LDA	HDP-LDA
11 月 15 日	10	9	8
実行時間 (sec)	890.78	5852.84	926.15

表 2: トピック抽出結果 (term-score 上位単語)

	トピック	term-score 上位単語	ラベル	カテゴリ
15日	topic0	逮捕 航海 捜査 取り調べ 方針	航海士の懲罰	映像流出
	topic1	捜査 映像 神戸 海保 中国	映像の管理	映像流出
	topic2	選挙 事務所 広報 政策 見直し	福岡市長候補者の訴え	その他
	topic3	大使館 中国 郵送 金属 ライフル	中国への批判	映像流出
	topic4	中国 合意 外相 再開 前原	日中外相会談	APEC
	topic5	映像 航海 投稿 削除 私用	流出後の映像の扱い	映像流出
	topic6	沖 集合 時間 政府 日本	日本への批判	映像流出
	topic7	高島 福岡 吉田 自民 民主 政権	福岡市長選結果	その他
	topic8	公開 衆院 与党 自民党 審議	不信任案の可決	映像流出
	topic9	首脳 会談 会見 政府 領土	日本政府の会談への見解	APEC

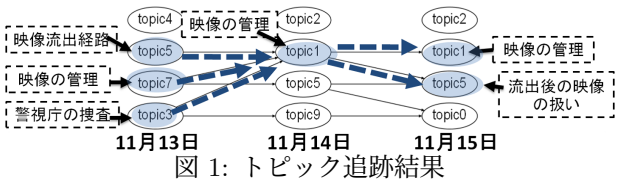

図 1: トピック追跡結果

トピック数については，手法間で大きな違いは見られない．また，実行時間も，HDP-LDA とほぼ同等であり，提案手法は HDP-LDA に敵う結果をもたらすことが分かった．実際に抽出されたトピックについても，表 2 から，その日の話題が重複することなく抽出されており，適切なトピック抽出が出来たと解釈できる．

また，トピックの追跡については,「映像流出問題」のうち「流出した映像」に関する話題のみが追跡されているなど，関連する話題のみが追跡できていることが分かる．

3　潜在的意味を考慮した適合フィードバック

3.1　概要

適合フィードバック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クエリ拡張手法の一つである．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う．

3.2 手法

まず，単語出現確率の最尤推定値に Dirichlet 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文書を対象に KL ダイバージェンス検索モデル (KLD)[2] 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．次に，LDA によりクエリと文書群 D のトピック分布を推定し，それぞれトピックベースのモデル P_q，P_D へ変換する．トピックベースのモデルでは (1) トピック分布 θ_d，(2) 推定された単語出現確率 $\sum_{j=1}^{T} \theta_{d,j}\phi_{j,w}$ を用いて 2 通りに文書を表現する．また，D の内上位 n 件を擬似関連文書と見なし，関連文書の各素性の重みの平均をフィードバックモデル P_F とする．最後に，パラメータ $a(0 \leq a \leq 1)$ を用いて以下の式によりクエリを更新する．

$$P_{q'} = (1-a)P_q + aP_F \qquad (2)$$

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

3.3 実験

3.3.1 実験仕様

NTCIR-2 の情報検索システム評価用テストコレクションの日本語検索課題 30 件を用い，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 $m = 100$，擬似関連文書と見なす文書数 $n = 10$ とした．実験では (a) クエリ更新式 (2) 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．

3.3.2 結果と考察

トピックを考慮した手法（トピックベース手法）とフィードバックにトピック情報でなく初期検索と同様のユニグラムモデルで文書を表現した表層情報のみの手法（表層ベース手法）それぞれの P@10 評価結果を図 2 と図 3 に示す．図 2,3 の各線は初期検索精度毎の P@10 の値を示している．

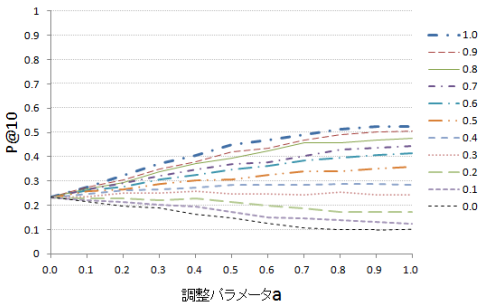

図 2: トピックベース手法の結果

初期検索精度が 0.0〜0.2 と低い場合は，表層ベース手法に比べ，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

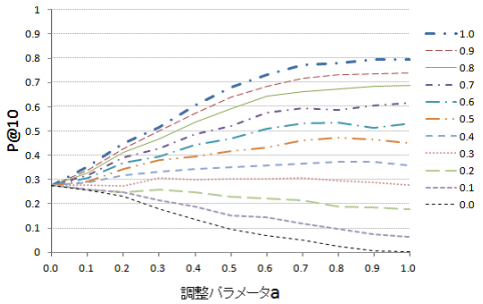

図 3: 表層ベース手法の結果

を上回っている．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ことが分かった．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はトピックを考慮した手法は有効であると考えられる．

4 おわりに

本研究では，文書の持つ潜在トピックの関連性に着目し，トピック追跡と適合フィードバックへのトピックモデルの適用手法を提案し，潜在的意味の有効性を考察した．前者では，トピック間類似度を用いたトピック抽出と追跡を行い，実験により適切なトピックの抽出が出来ることを示した．後者では，適合フィードバック手法にトピックモデルを導入した手法を提案し，実験によりフィードバックへのユーザの要求の反映が不十分な場合にはトピックを考慮した手法は有効に働くことが分かった．今後の課題としては，現在，いずれの手法においてもパラメータに依存してしまう問題点があるため，各変数についての検証などが挙げられる．

参考文献

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

[2] 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.

[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.