4B4-5

facet-biased トピックモデルと距離尺度学習を用いたニュース記事の分類

Classification of news article by using facet-biased topic model and distance metric learning

小野寺 大輝 *1　黄 楽 *1　吉岡 真治 *1
Daiki ONODERA　HUANG Le　Masaharu YOSHIOKA

*1北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

Nowadays, we can access varieties of news articles from different news sites. However, due to the large numbers of accessible news articles, news aggregation sites or recommendation sites that classify news articles are used for finding out interesting ones. In this paper, we propose a framework to classify news articles based on the facet-biased topic model and distance metric learning. Facet-biased topic model is an extension of topic model that construct topics for particular facets in addition to the ordinal general topics. Distance metric constructed by this topic model is modified by using Large Margin Nearest Neighbor; distance metric learning method for k-NN classification. The perfomance of the proposed framework is evaluated by using the crawled data from Yahoo! News and confirmed effectiveness of facet-biased topic model and distance metric learning.

1. はじめに

近年、新聞社やテレビ局といった報道機関に加え、様々なニュースサイトから、多くのニュース記事が配信されている。個々のユーザにとって、これらのニュース記事の中から興味のあるニュース記事を見つける事は困難であるため、Yahoo! News*1 や Google news*2 といったニュースアグリゲーションサイトやニュース推薦のアプリ (smartnews*3, グノシー *4) などが利用されている。これらのサイトでは、まず、ニュース記事を特定のカテゴリに分類して整理し、その中からユーザが興味を持つであろう記事の推薦などを行っている。本研究では、このニュース記事の分類という作業に注目し、その作業を支援するための基本技術としてのニュース記事を分類する枠組の提案を行う。本研究では、記事の分類の枠組として、距離尺度学習の枠組として Large Margin Nearest Neighbor[Kilian Q. Weinberger 09]。を用いた k-Nearest Neighbor(kNN) を利用する。また、もとのニュース記事を Bag-of-Words の高次元の空間からの圧縮を行うために、facet-biased トピックモデルを利用する。facet-biased トピックモデルは、既存のトピックモデルの枠組において、特定の facet(観点) とその関連語を生成するためのトピックという考え方を導入したトピックモデルである。本研究では、ニュース記事の分類においては、芸能人やチーム名といった固有名や、日本の地名か外国の地名などの存在が重要な役割を果たすことに注目し、人名・地名・組織名という facet を考え、全体のトピック群が人名とその関連語を生成するトピック、地名とその関連語を生成するトピック、組織名とその関連語を生成するトピック、その他の語を生成するトピックから構成されると仮定したモデルを生成する。また、Yahoo! News からクロールしたデータを用いて分類実験を行い、本システムの有効性を検討する。

連絡先: 小野寺 大輝, 北海道大学大学院情報科学研究科, 札幌市北区北 14 条西 9 丁目,011-706-7161, onodera@kb.ist.hokudai.ac.jp

∗1 http://news.yahoo.co.jp/
∗2 https://news.google.co.jp/
∗3 https://www.smartnews.com/
∗4 https://gunosy.com/

2. facet-biased トピックモデルと距離尺度学習

2.1 facet-biased トピックモデル

トピックモデルとは、単語の生成確率により特徴づけられるトピック (例えば、スポーツに関係するトピックならば、スポーツに関連するトピックの単語の生成確率は高く、その他の単語の生成確率は低い) の組合わせにより、各々の文書が生成されると考える文書の生成モデルであり、代表的な手法としては LDA[Blei 03] がある。

LDA では単語間に現れる単語間の共起情報を用い、全体の文書群中で特徴的に現れるような共起語群を中心としたトピックを生成し、各々の文書におけるトピックの混合比率を推定することにより、各々の文書の特徴を表すことが可能である。

facet-biased トピックモデル [小野寺 16] は、このトピックモデルを拡張し、トピックモデルが生成するトピックに、特定の facet(観点) とその関連語を生成するためのトピックという考え方を導入したトピックモデルである。具体的には、特許文書に対して、対象 (IC チップなどの対象語により表現) と観点 (コスト、信頼性などの観点語により表現) という二つの facet を考え、各特許文書が、二つの facet を代表するトピックと、その他のトピックという、3 つのタイプのトピックの混合として生成されると考えたトピックモデルである。

このトピックモデルでは、facet を代表するトピックが facet 語とその facet 語に関連するその他の語のみを生成するという制約を与えることにより、facet を代表するトピックの生成を行う。この結果、全体の文書中に存在する各々の facet 語が、その他の語の中に存在する特徴的な共起語を含む形でトピックに割り当てられ、各々の facet 語との共起性を考慮した分類が行われる。

この facet-biased トピックモデルを用いることにより、特許文書中の全ての文書について対象語と観点語についてトピックが割り当てられると共に、各々の単語についても対象語と観点語についてのトピックが割り当てられ、結果として、これらのトピックの情報を用いて特許文書を特許マップに配置したり、特許文書中の対象語と観点語を含むような文単位で特許マップ上に配置したりすることが可能になる (図 1)。

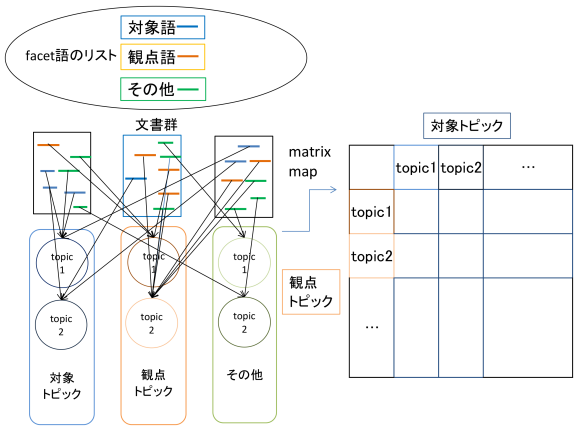

図 1: facet-biased トピックモデルによる特許マップの作成

2.2 Large Margin Nearest Neighbor

k-Nearest Neighbors (kNN) は、分類問題を解くための良く知られたアルゴリズムである。このアルゴリズムでは、既に分類ラベルが付与されているデータが配置されている空間内において、分類対象とする新しいデータの k 個の最近傍のデータのラベルのうち、最も一般的なラベルをそのデータに割り当てることによって、新しいデータに対する分類を与える方法である。本手法では、近傍データを見つけるための距離が非常に重要な役割を果たす。一般に、多次元空間における距離尺度として用いられるものとして、ユークリッド距離などがあるが、必ずしも、一般的なユークリッド距離が適切な分類に役立つ保証はない。

この様な問題を解決するために、Weinberger ら [Kilian Q. Weinberger 09] は、この kNN の分類性能を向上させるための距離尺度の学習アルゴリズムを提案している。このアルゴリズムでは、既存の与えられた分類ラベルの情報を用いて、以下の 2 つの基準を可能な限り満足させる距離尺度を学習し、その距離尺度を用いた分類を行うことで、既存のユークリッド距離を用いた分類よりも分類性能の向上をはかる。

Push k 個の最近傍のデータが、出来る限り同じラベルに属する。

Pull 異なるクラスに属するデータが近くに存在する場合には、同じラベルのデータよりも与えられたマージン以上の距離だけ遠ざける。

つまり、Push の基準により、同じラベルに属するデータを近くに配置し、Pull の基準により、異なるラベルに属するデータを遠くに配置することのよって、kNN の分類誤りが少ない距離尺度を得ようという考え方である。LMNN では、上記の基準を満たすように最適化された距離空間を、元の多次元空間からの線型変換によって作成するための線型変換行列 L を求める問題として定式化し、i 番目と j 番目のデータに対応するベクトル x_i,x_j の距離を以下の式によって計算する。

$$D(\vec{x_i}, \vec{x_j}) = ||L(\vec{x_i} - \vec{x_j})|| \quad (1)$$

この行列 L を求める問題を、push と pull の基準に対応する次のようなコスト関数をの最小化問題として解く。

$$\epsilon_{pull}(L) = \sum_{j \rightsquigarrow i} ||L(\vec{x_i} - \vec{x_j})||^2 \quad (2)$$

$$\epsilon_{push}(L) = \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{il}) \left[1 + ||L(\vec{x_i} - \vec{x_j})||^2 - ||L(\vec{x_i} - \vec{x_l})||^2\right]_+ \quad (3)$$

$$\epsilon(L) = (1 - \mu)\epsilon_{pull}(L) + \mu\epsilon_{push}(L) \quad (4)$$

ただし、

- y_{il} は x_i と x_j が同じラベルに属するときのみ $y_{il} = 1$ とし、そうでない場合は、$y_{il} = 0$ とする。
- $[z]_+ = max(z, 0)$ は、一般的なヒンジロス関数である。
- $j \rightsquigarrow i$ は $\vec{x_j}$ が $\vec{x_i}$ にとっての k 最近傍であることを示す。

3. facet-biased トピックモデルと距離尺度学習を用いたニュース記事の分類

3.1 ニュース記事分類のための facet-biased トピックモデル

ニュースの記事分類では、固有名の存在が大きな役割を果たす。例えば、野球の始球式のニュース記事であっても、投げた人が芸能人であれば、芸能のトピックとして扱われ、アメリカ大統領が投げれば国際や政治の記事となる。また、地震に関する記事においても、日本で起きた地震は国内の記事になるが、海外で起きた地震は海外の記事となる。このような記事を分類するためには、一般的な語についての特徴を見るだけでなく、固有名に注目した特徴を考えることが有用であると考えた。

トピックモデルに代表されるような一般的な次元圧縮の手法では、このような固有名に注目した特徴を明示的に扱う方法は存在しない。そこで、本研究では、facet-biased トピックモデル [小野寺 16] の facet として、人名、地名、組織名という 3 つの facet を考え、これらの facet に関連するトピックを作成することにより、特徴的な固有名の共起性に基づくトピックが生成する。このような固有名に注目したトピックの混合比率として文書を表現することにより、固有名のタイプ (トピックとして表現される) を考慮した文書の特徴量が生成可能となる。

従来の facet-biased トピックモデルでは、facet として 2 つの facet が利用されていたが、今回は 3 つの facet を利用する。この facet-biased トピックモデルでは、facet に関連するトピックでは、facet に属する語と facet に関連する語のみが生成されるという考えから、表 1 で示すような facet に関連するトピックとその単語の発生確率を用いてトピックの推定を行う。ここで、0 は、トピックモデルの枠組で与えられる正則化のための 0 に近い確率を与えることとする。

表 1: 単語の発生確率の設定

	人名	組織名	地名	その他の語
人名 (facet1)	推定値	0	0	推定値
組織名 (facet2)	0	推定値	0	推定値
地名 (facet3)	0	0	推定値	推定値
その他	0	0	0	推定値

3.2 距離尺度学習を用いたニュース記事分類

前節で作成した facet-biased トピックモデルを用いることにより、固有名に関連するトピックが生成されることになるが、これらの全ての固有名のトピックが等しく分類に寄与するとは限らない。よって、本研究では、前節で作成した facet-biased トピックモデルにより得られた各文書のトピックの混合比率によって生成される次元圧縮された空間上の情報を LMNN を用いて距離尺度学習を行うことにより、より分類に役立つような距離尺度を作成し、実際の分類を行う。

4. 実験と考察

本研究で提案する facet-biased トピックモデルと距離尺度学習を用いたニュース記事の分類の枠組の有用性を検討するために、実際のニュース記事を用いた分類実験を行う。

具体的には、Yahoo! News で公開されているニュース記事を対象に、Yahoo! News が分類した記事分類を正解データとして、その正解データをどれくらい再現できるかという実験を行った。具体的には、2015 年 11 月 1 日から、2016 年 1 月 14 日にクロールしたデータを用い、4 週間分の記事を学習データ (平均記事数 62,648.6 件) として、次の 1 週間の記事分類を行うというテストデータ (平均記事数 14,642.4 件) を 5 組 (学習データの先頭の日付が、11/1,11,21,12/1,11) 作成し、各々の組について、2 つのトピックモデル (通常のトピックモデルもしくは facet-biased トピックモデル)、距離尺度学習の有無、2 つのトピック数 (t=100,t=200) の組み合わせによる計 8 つの設定により実験を行った。

トピックモデルについては、scikit-learn[*5] の online batch 処理によるトピックモデルのライブラリ [*6] を利用し、facet-biased トピックモデルについては、表 1 に示す単語の発生確率の制約を満すように、上記のライブラリを修正したものを利用した。また、トピックモデルに関連するパラメータについては、システムのデフォルト設定を利用した。

facet-biased トピックモデルで用いる人名・組織名・地名の推定については、これまでの研究 [吉岡 11] で利用してきた Wikipedia のエントリーに基づく辞書で拡張をした MeCab と Cabocha を用いた固有名抽出システムを利用して、その推定を行った。また、その他の語としては、代名詞、接尾、数などを除く一般名詞を利用した。通常のトピックモデルを用いる場合においても、固有名も利用して文書ベクトルの作成を行った。

各々の文書に対応する文書ベクトルの作成時には、全体の文書群で 3 回以下しか現れない低頻度語を削除すると共に、文書中に存在する語に対応する重みとしては、TF-IDF を利用した。また、文書長を考慮した正規化を行うために、総和を 1 に正規化したベクトルを利用して、トピックモデルの生成を行った。

また、facet-biased トピックモデルの生成時には、各々の facet に関係するトピックの数を設定する必要がある。特許マップの生成時 [小野寺 16] には、10x10 の特許マップの生成を目標として、各 facet に対応するトピック、その他のトピックを同じく 10 としていた。しかし、本研究では、これらのトピック数を同じにすると、固有名に関するトピックが多数を占めることになり、その他の語に関するトピックが十分に作れない可能性が高くなる。よって、本研究では、先の文書ベクトルの作成時に用いた各 facet 語の異なり語数に対応する形で、全体のトピックを分割することとした。例えば、人名:1000 語、地名：500 語、組織名：500 語、その他：3000 語で 100 トピックの場合は、人名：20 トピック, 地名：10 トピック, 組織名:10 トピック, その他:60 トピックと設定した。

トピックモデルにおける各文書の混合比率には、トピックに対応する語が存在しない場合においても、正則化のための非常に小さな定数比率を与えているが、本研究で考えているような文書分類の観点からは、この様な定数には意味がないと考え、これらの値を 0 として扱うこととした。また、先ほどのライブラリが出力する混合比率については、その大きさが正規化されていないため、先ほどの文書ベクトルの時と同じく、総和を 1 にする正規化を行った。

この結果得られたベクトルに対し、距離尺度学習を行わない場合は、そのまま kNN による分類を行い、距離尺度学習を行う場合には、LMNN[Kilian Q. Weinberger 09] の実装である LMNN のライブラリ [*7] を用いて距離尺度学習を行いその結果を用いて kNN による分類を行う。ただし、LMNN の計算コストが非常に高いため、7000 記事 (約 10%) のランダムサブサンプルを用いた計算を行った。また、kNN の k としては、5 を利用し、それ以外のパラメータはライブラリのデフォルト設定を利用した。

これらの設定を用いて、5 つのテストデータについて分類実験を行った。表 2 に 8 つの設定についての分類精度に関する平均を示す。

表 2: 各手法による 1 週間の記事の分類精度

	facet-biased		通常のトピックモデル	
	LMNN	LMNN なし	LMNN	LMNN なし
t=100	0.763	0.758	0.717	0.714
t=200	0.764	0.754	0.714	0.715

個別の設定ごとの性能を分析するには、事例 (5) が少ないこともあり、全ての実験について、他の設定の組み合わせ (2x2=4) を考慮した 20 の対応するデータについて、ウィルコクソンの符合つき順位和検定により、トピックモデルの種類、LMNN の利用、トピック数について、有意性の分析を行った。トピックモデルについては、全ての組合わせについて一貫して、通常のトピックモデルより分類精度が高く、$p < 0.01$ で有意性が確認された。また、LMNN については、通常のトピックモデルの 200 トピックの場合をのぞいた 3 種類の組合わせにおいては、LMNN を利用した方が一貫して分類精度が高く、同じく、$p < 0.01$ で有意性が確認された。ただし、通常のトピックモデルを用いた場合には、その改善の幅は小さかった。トピック数については、100 と 200 の差はほとんどなく、また、有意性も確認できなかった。

この事から、facet-biased トピックモデルで生成したトピックの方が通常のトピックモデルで生成するトピックよりも、全体の文書を分類するという観点からは性能が良く、さらに、LMNN による線型変換を行うための基底としても良い性能を持っていると考えられる。

次に、記事の分類ごとでの分類性能を議論する。表 3 に 2015 年 11 月 1 日のデータについて、提案手法 (200 トピックの facet-biased トピックモデルと距離尺度学習を利用) による分類毎の分類結果 (精度と再現率) を示す。

表 3: 2015/11/1 のデータに対する提案手法の分類毎の記事数と精度

分類名	記事数	正解記事数	精度	再現率
スポーツ	4,339	4,113	0.950	0.948
エンタメ	3,633	3,313	0.841	0.942
国際	2,129	1,353	0.720	0.636
国内	1,961	1,079	0.612	0.550
経済	1,763	1,215	0.621	0.690
地域	1,109	390	0.470	0.352
IT ・科学	888	551	0.594	0.620
ライフ	523	204	0.284	0.390
全体	16,345	12,218	0.748(分類精度)	

この表から分かるように、テストデータには、分類ごとに、記事数の大きな片寄りが存在する。これは、学習データについ

*5 http://scikit-learn.org/

*6 http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

*7 LMNN version 3 https://bitbucket.org/mlcircus/lmnn/

ても同じ傾向がある。また、記事数の大きなスポーツや芸能の分類性能は良く、記事数の少ないライフの分類性能が悪い。これは、記事数のアンバランスのために、スポーツや芸能に関する記事は十分多く存在するため、これらの記事に関するトピックは生成されやすく、記事数の少ない分類に関するトピックは、ほとんど作られずに結果として、分類に役立つトピックが少なくなってしまっていることが原因ではないかと考えた。

この問題を解決するために、記事数の大きな分類から順番に記事分類を推定していく逐次的な分類を行うアルゴリズムを作成した。具体的には、次のような手順で分類を行う。

1. 記事数の多さを考慮して、分類の順序を「スポーツ → エンタメ → 国際 → 国内 → 経済 → 地域 →IT ・科学 → ライフ」のように設定し、最初は、全ての学習データを対象学習記事、全てのテストデータを対象テスト記事、「スポーツ」を対象分類と設定する。
2. 対象学習記事を全て利用して、設定に応じたトピックモデルの学習、LMNN による学習 (7,000 記事より多い場合には、7000 記事のサブサンプルを利用) を行い、kNN の分類器を作成する。この分類器を用いて、全ての対象テスト記事の分類を行い、対象分類と判定された記事の分類のみを確定する。ただし、分類対象が 2 つの場合には、全ての分類を確定し、最終結果とする。
3. 対象学習記事から、対象分類に属する記事を取り除いたものを新たな対象学習記事とし、分類の確定しなかった記事群を次の段階の対象テスト記事とする。分類の順序に応じて、次の分類 (「スポーツ」の次は「エンタメ」) を対象分類と設定し、2. の手順に戻る。

この様な手順を繰り返すことにより、記事数の少ない分類に対しても、その分類を確定する際には、分類に関連するトピックが十分存在する中で分類が行われることが期待される。

このアルゴリズムによって、生成された結果を表 4 に示す。

表 4: 各手法による 1 週間の記事の分類精度

	facet-biased		通常のトピックモデル	
	LMNN	LMNN なし	LMNN	LMNN なし
t=100	0.778	0.772	0.734	0.734
t=200	0.779	0.770	0.735	0.734

この逐次的な分類においても、トピックモデルの違い、LMNN の利用、トピック数についての傾向は変わらなかった。逐次的な分類と 1 回の分類の結果の比較を、実験回数、他の設定の組み合わせ (2x2x2=8) を考慮した 40 の対応するデータについて、ウィルコクソンの符合つき順位和検定により、その有意性の分析を行った。40 の組合わせの内、通常のトピックモデルを使ったトピック数 100 の場合のデータにおいて、LMNN ありの場合となしの場合の 2 回だけ逐次的な分類の方が性能が悪い場合があるだけであり、$p < 0.01$ で有意性が確認された。

次に、記事の分類ごとでの分類性能についてであるが、表 3 に対応する評価結果を表 5 に示す。下線部が、表 3 の結果より改善した部分となる。全体としては、「エンタメ」、「国際」、「国内」、「経済」といったスポーツの記事より記事数の少しだけ少ない分類での再現率が向上しており、結果として、全体の分類性能が向上していることが確認できた。

次に、記事数の少ないところ分類性能について議論する。IT ・科学とライフの記事 1,411 記事を対象として、最終的に作成した逐次的分類の分類システム (IT・科学とライフの 2 クラスに分類する) を用いて分類した結果は、分類精度が 0.865(1,221/1,411) であった。よって、逐次的に作成した分類システムは、分類システムとしては、他の分類に比べても悪い

表 5: 2015/11/1 のデータに対する逐次分類を行う提案手法の分類毎の記事数と精度

分類名	記事数	正解記事数	精度	再現率
スポーツ	4,339	**4,113**	0.950	0.948
エンタメ	3,633	**3,445**	0.812	**0.948**
国際	2,129	**1,574**	0.689	**0.740**
国内	1,961	**1,206**	0.605	**0.615**
経済	1,763	**1,238**	**0.675**	**0.702**
地域	1,109	369	**0.524**	0.333
IT ・科学	888	504	**0.746**	0.568
ライフ	523	131	**0.470**	0.250
全体	16,345	**12,580**	**0.770(分類精度)**	

システムではない。しかし、本来正解と判断できる記事が逐次的な処理の中で別の分類に割り当てられたり、どちらのラベルをつけても正解とならないような、他の分類で分類に失敗した記事などの存在が、この様な結果になった原因だと考えられる。

5. おわりに

本研究では、facet-biased トピックモデルと距離尺度学習を用いたニュース記事の分類手法の提案を行った。実験結果から、ニュース記事の分類という観点からは facet-biased トピックモデルは、トピックモデルよりも有意に性能が良いことが確認された。また、距離尺度学習を行うことにより、僅かではあるが、有意に性能が向上することも確認した。さらに、分類語との記事数に片寄りがある場合に記事数の少ない分類に関する有用なトピックが作られないということを考慮した逐次的な分類モデルを用いることにより、有意な精度の向上が出来ることについても確認した。

しかし、現時点での分類精度は、80%を下回る状況であり、単独のシステムとしてニュース分類を行うには、まだまだ不十分である。全体の分類精度の向上のためには、トレーニングデータに用いる記事数や facet-biased トピックモデルに関する設定も含め、さらなるパラメータに関する検討を行うと共に、ここで得られた情報に加え、記事の発信元、発信元で付与された分類の情報なども考慮した総合的なシステムの構築が必要であると考えている。

謝辞

また、本研究の一部は、科研費基盤研究 (B) 25280035 により行われた。ここに記して、謝意をあらわす。

参考文献

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *the Journal of machine Learning research*, Vol. 3, pp. 993–1022 (2003)

[Kilian Q. Weinberger 09] Kilian Q. Weinberger, L. K. S.: Distance Metric Learning for Large Margin Nearest Neighbor Classification, *Journal of Machine Learning Research*, Vol. 10, pp. 207–244 (2009)

[吉岡 11] 吉岡 真治, 神門 典子, 関 洋平 : 複数国の新聞サイトを比較分析する NSContrast の実験的分析, 情報処理学会デジタルドキュメント研究会, 2011-IFAT-103 (2011), IFAT-103-2

[小野寺 16] 小野寺 大輝, 吉岡 真治 : 対象-観点を考慮した facet-biased トピックモデルと特許マップへの応用, 言語処理学会第 22 回年次大会発表論文集 (2016), P13-5