3N-1

Information Bottleneck法の文書クラスタリングへの適用

矢口 輝和　　田中 譲
北海道大学大学院情報科学研究科
コンピュータサイエンス専攻知識メディア研究室

1　はじめに

現在，単純な文書や画像，音声・動画にとどまらず，書籍や百科事典，遺伝子データなどと言った様々な情報の電子化・データベース化が進んでいる．また，インターネットの普及，ブロードバンド化によって，それらの情報に簡単にアクセス共有できるようになった．特に文書に関して言えば，ブログが爆発的に普及しており，我々が利用できる情報は膨大である．

本研究では，こうした現状をふまえ，大量の情報から効率的に必要な情報を抽出すべく，Information Bottleneck 法に基づいた，高精度な文書クラスタリング技術の実現を目指す．この実現のため，同一単語であってもコーパス中の出現位置によって区別し，話題の転換点として段落を考慮するという手法を提案する．同理論によってパラメータの入力や辞書などが不要となり，さらに提案手法によって，ある単語が複数の意味で使用される，ある文書が複数の話題について言及しているといった曖昧性を除去する．教師なしの手法でありながら，他の教師あり手法の精度を上回ること，従来手法の欠点であるクラスタ数が多くなった場合の精度低下を軽減することが目標である．

2　Information Bottleneck

Information Bottleneck(以下 IB) は N.Tishby ら [2] が 1999 年に提案した情報理論に基づく手法である (図 2)．本研究のテーマである文書クラスタリングでは，圧縮される文書集合が X，クラスタリングの際の指標となるもの，例えば文書に含まれる単語集合 (トピック) が Y(relevant variable) であり，X のクラスタリング結果が T である．IB では 2 つの相互情報量によって結果 T の圧縮率 $I(X;T)$，正確性 $I(T;Y)$ を評価し，次式の最大化を図る．

$$\mathcal{L}_{max} = I(T;Y) - \beta^{-1} I(T;X)$$
$$\beta \in [0, \infty) : \text{ラグランジュ乗数}$$

つまり，X と T の間の相互情報量 $I(X;T)$ を最小化する一方で，T と Y の間の相互情報量 $I(T;Y)$ を最大化しようとし，両者のトレード・オフによって問題の最適化を図る手法である．IB は指標としての変数 Y を導入することによって，ユーザはクラスタ数のみ設定すれば実行可能となった．しきい値などのパラメータを設定することに比べてはるかに容易であることは明白であり，結果に非常に影響力のある量をデータの外部に求めないことの意味は大きい．

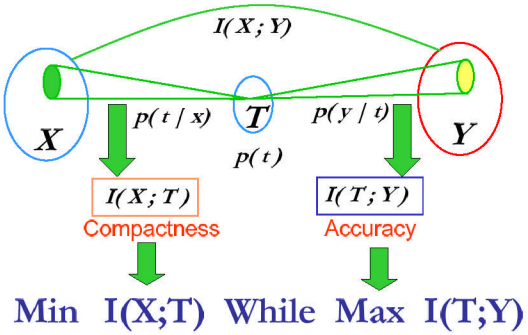

図 1: Information Bottleneck 概念図

具体的なクラスタリングアルゴリズムについても N.Tishby らによっていくつか紹介されているが，本研究ではその中で精度・処理時間等に優れた sIB(sequential optimization) アルゴリズムを用いている．sIB アルゴリズムは”sequential K-means like”なアルゴリズムで，平面的なクラスタリングを行うものである．

[1] で報告されているクラスタリング精度の比較結果を見ると，IB(sIB) は教師なし手法である K-means アルゴリズムを大きく上回り，教師つき手法である Naive Bayes(NB) アルゴリズムに迫る精度を示している．しかし，どの手法にもいえることであるが，クラスタ数が多くなるほど精度は低下してしまう．

3　曖昧性の低減

3.1　2 つの曖昧性

我々は，この原因を次の 2 つの曖昧性によるものと考えている．

(a) ある単語が複数の意味を持つ

(b) ある文書が複数のトピックについて言及している

Applying Information Bottleneck to Document Clustering.
Terukazu Yaguchi , Yuzuru Tanaka
Meme Media Laboratory,Hokkaido University
N13W8,kita-ku,Sapporo,060 8628,Japan

文書クラスタリングを行なう際，一般的な手法では各文書に対して単語の生起回数をカウントし，その共起関係を見る．ここで我々が注目したのは，表記は同じでも複数の意味を持つ単語が非常に多いということである．同一表記の単語であるからと一様にカウントしていたのでは，こういった違いが考慮されず，本来区別されるべき文書同士を関連付けてしまい，精度を低下させる可能性がある．また，ある単語が様々なトピックの中で用いられる場合も同様である．複数のトピックに言及する文書が，トピックの異なる文書同士を関連付ける橋渡しをしている可能性もある．よって，各文書の主題が何であるのかを考慮する必要がある．

3.2 単語の出現位置，段落の考慮

我々は，異なる段落に出現する単語は同一表記のものであっても区別することで，これらの曖昧性の除去を図る．これは，単語の意味を決定するのはその単語の出現する文書・文脈であり，各段落は1つのトピックについて述べているという仮定に基づいている．まず，文書クラスタリング (X：文書，Y：単語) の前処理として，同一表記の単語ごとに段落の観点からクラスタリングを行う．つまり，図3のように X が単語 (例：$\mathrm{dog}_1, \ldots, \mathrm{dog}_n$)，$Y$ を段落として，表記ではなく，意味の観点から，より詳細に単語を分析する．この処理で得られた結果 T(例：$\mathrm{dog}_1, \ldots, \mathrm{dog}_{n'}, \mathrm{cat}_1, \ldots, \mathrm{cat}_{n'}, \ldots$) を最終的な文書クラスタリングの Y とすることで，判断指標をより明確にする．しかし，各単語をクラスタリングする際，それぞれがいくつの意味 (クラスタ数) を持つのかを動的に決定することが望ましいが，実際には困難である．必要以上に大きくすると，似通った文書同士の関連性をも低くしてしまうので逆効果であるから，その決定手法は非常に重要な要素となる．我々はその手法についても各単語と段落の関連において定義できるのではないかと考え，次のような処理を行った．

	X	Y	T
(1)	P	W	$\widetilde{P}^{(0)}$
(2)	W	$\widetilde{P}^{(i)}$	$\widetilde{W}^{(i)}$
(3)	P	$\widetilde{W}^{(i)}$	$\widetilde{P}^{(i+1)}$
(4)	D	$\widetilde{W}^{(i)}$	$\widetilde{D}$

W :段落ごとに区別された単語の集合
P :段落
D :文書

図 2: 処理手順

(2)(3) のクラスタリングを繰り返し行い，前後の結果を比較する過程で決定できないかと試みている．2つのトピックから集めた文書に対してクラスタ数を2に固定して処理した結果，繰り返しの前後で各単語に割り振られるクラスタ番号にほぼ変化はなかった ((2) の $\widetilde{W}^{(i)}$)．このとき，文書クラスタリングの結果，Recall ratio90%以上を達成した．これに対し，10のトピックから集めた文書に対して処理した結果，20%程度の変動が見られ，文書クラスタリングの結果も Recall ratio50%程度にとどまった．つまり，単語をより詳細に定義し，前後の変動を低く抑えられれば，より正確なクラスタリングが実現できると考える．

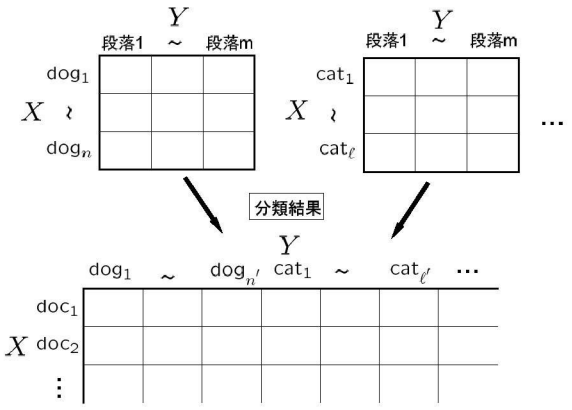

図 3: 共起行列の作成

4 おわりに

IB は既存手法に比較して高い精度を示すことがすでに報告されており，多くの分野への適用が試みられている．我々は更なる精度向上ために，特にクラスタ数の大きい場合について問題となりうる2つの曖昧性を除去するため，単語の出現位置や段落単位で文書を分割するという手法を提案した．これによって各単語の意味をより厳密に定義できれば，精度向上へつながるのではないかと考えている．

参考文献

[1] Noam Slonim. The Information Bottleneck: Theory and Applications. Doctors Thesis, School of Computer Science and Engineering, Hebrew university, 2002.

[2] N.Tishby, F.Pereira, W.Bialek. The Information Bottleneck Method. Proc. 37th Allerton Conference on Communication and Computation, 1999.