テキスト分類のための潜在トピックを考慮したグラフ構成

Latent Topic-based Graph Construction for Text Classification

江里口 瑛子 [1] *　　小林 一郎 [1]
Akiko Eriguchi[1]　　Ichiro Kobayashi[1]

[1] 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
[1] Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Abstract: This paper aims to raise the accuracy of multi-class text classification by means of graph-based semi-supervised learning (GBSSL). It is essential to construct a proper graph expressing the relation among nodes in GBSSL. We propose a method to construct a similarity graph by employing both surface information and latent information to express similarity between nodes. Experimenting on Reuters-21578 corpus, we have confirmed that our proposed method works well for raising the accuracy of GBSSL in multi-class text classification task.

1　序論

機械学習手法は，教師あり学習，教師なし学習，半教師あり学習などがある．半教師あり学習 (Semi-Supervised Learning: SSL) 法は，少量のラベルありデータを用いて，多量のラベルなしデータに付与するラベルを予測する手法である．その中でも，グラフ構造に基づく半教師あり学習 (Graph-Based Semi-Supervised Learning: GBSSL) 法は，文書分類タスクにおいて，Support Vector Machine (SVM)[2] などの学習法と比べてより有効な手法であることが知られている [4].

GBSSL 法の精度は，一方でどのような教師データ (ラベルありデータ) を与えるかによって左右され，他方で，どのようなグラフを構成するかによって左右されることが分かっている [7, 9]．前者に関連して重要となるのが，どのようにして情報量の大きいデータを選出するかである．その良い事例が能動学習法であり，質の高い教師データを選出するための方法である．GBSSL 法の精度を改善するため，いくつかの能動学習法が提案されている [7, 10]．また，後者に関連して重要となるのは，グラフのノード間の関係性をどのように表現するかである [9]．一般に，GBSSL 法のグラフスパース化手法には，k-近傍グラフが用いられることが多い．しかしながら，k-近傍グラフではその構成上，ハブ点と呼ばれる高次数のノードができやすく，このハブ点は GBSSL 法の精度を悪化させるということが報告されている [11]．ノードに次数制約を設けた，グラフスパース化手法もまたいくつか提案されている [11, 12].

本研究では，GBSSL 法を用いた多クラス文書分類におけるグラフ構成手法の提案を行う．グラフ構成において，必須の要件であるノード間の類似度に，文書間の潜在的な類似度を新たに取り入れる．一般にこれまで，テキストデータから構成されるグラフにおいては，単語の頻度情報に基づく文書間の表層的な類似度が多く採用されてきたが，我々はこれに加えて新たに，確率的言語モデルに基づく文書間の潜在的な類似度を加えたものをノード間の類似度として採用する．また，これら表層的な類似度と潜在的な類似度を $(1-\alpha):\alpha$ $(0 \leq \alpha \leq 1)$ の割合で混合させ，α をパラメータとして動かし，両情報を同時に採用する．

上記手法をマルチラベルを有するテキストのカテゴリ分類に適用し，精度 PRBEP を算出し，我々の手法の有効性を各カテゴリ毎に評価し，かつ，それら全体の精度の向上を検討する．

2　文書分類のための GBSSL 手法

本研究で提示する，多クラス文書分類のタスクにおける GBSSL 法の詳細は，以下に述べる通りである．

2.1　グラフ構成

本研究におけるグラフ構成は，テキストデータを対象にして行う．したがって，各文書はグラフのノードとみなされる．そのノード (文書) 間の関係は類似度として表され，その類似度をグラフの辺の重みとするような重み付き無向グラフ $G = (V, E)$ を構成する．ここ

*連絡先： 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情報科学コース
〒 112-8610 東京都文京区大塚 2-1-1
E-mail: g0920506@is.ocha.ac.jp

で V と E は，それぞれグラフのノード集合と辺集合を表す．

グラフ G は隣接行列 $\boldsymbol{W}$ の形で表現することができ，$w_{ij} \in \boldsymbol{W}$ はノード i，ノード j 間の類似度を表すとする．特に，GBSSL 法の場合には，その類似度はノード i の k-近傍点集合 $K(i)$ からなるものとし，$w_{ij} = sim(\boldsymbol{x}_i, \boldsymbol{x}_j) * \delta(j \in K(i))$ とする．ここで，$\delta(z)$ は z が真ならば 1，偽ならば 0 とする．

2.2 グラフにおける類似度

テキストデータにおける文書間の類似度を測る指標として，表層情報に基づく類似度と潜在情報に基づく類似度の二種類の類似度を採用する．文書の表層情報としては，文書に含まれる単語の出現頻度に着目した *tfidf* ベクトル [3] が多く用いられる．ここでは，表層情報に基づく類似度 ($sim_{surface}$) を，*tfidf* ベクトルのコサイン類似度の値とする．また，文書の潜在情報として，複数文書内に隠れトピックが存在することを仮定し，その隠れトピックに関して生起する単語の確率分布 (トピック分布) を用いる．ここでは，潜在情報に基づく類似度 (sim_{latent}) を，シグモイド関数 (式 (4))) を用いて，トピック分布間の距離を類似度に変換したものとする．トピック分布間の距離は $L2$ ノルム距離 (式 (5)) を用いて求める．トピック分布の推定には，Latent Dirichlet Allocation (LDA) 法 [1] を用いる．

本研究では，この従来の類似度 ($sim_{surface}$) に新たに，文書の持つ潜在情報に基づいた類似度 (sim_{latent}) を α $(0 \leq \alpha \leq 1)$ の割合で付加する．これら sim_{latent} と $sim_{surface}$ を $\alpha : (1-\alpha)$ $(0 \leq \alpha \leq 1)$ の割合で合算した値を，ノード間 (すなわち，文書 S と文書 T 間) の類似度 (sim_{nodes}) とする (式 (1))．P と Q は，それぞれ文書 S と文書 T に対するトピック分布を表す．

$$\begin{aligned} sim_{nodes}(S,T) &\equiv \alpha * sim_{latent}(P,Q) \\ &\quad +(1-\alpha) * sim_{surface}(S,T) \end{aligned} \quad (1)$$

$$sim_{surface}(S,T) = \cos(tfidf(S), tfidf(T)) \quad (2)$$

$$sim_{latent}(P,Q) = \frac{2}{1+\exp^{L^2(P,Q)}} \quad (3)$$

$$\sigma_1(x) = \frac{1}{1+\exp^{-x}} \quad (4)$$

$$L^2(P,Q) = \int (P(\boldsymbol{x}) - Q(\boldsymbol{x}))^2 d\boldsymbol{x} \quad (5)$$

2.3 ラベル伝搬法

本研究における GBSSL 法として，ラベル伝搬法 [5, 8] を採用する．ラベル伝搬法は，「グラフ上において，辺で繋がるノード同士は同じカテゴリに属す」という仮定に基づき，カテゴリラベル未知のノード (すなわち，テストデータ) について予測を行う手法である．

類似度行列を $\boldsymbol{W}$，ノード数を n 個 (このうち教師データ数は l 個) とする．n 個のノードに対する予測値 $\boldsymbol{f}$ は，以下の最適化問題の目的関数 (式 (6)) の解 (式 (8)) として求まる．式 (6) の第 1 項は，各ノードの予測値と教師データの正解値の差を表し，第 2 項は，類似度グラフ上で隣接するノード同士の予測値の差を表す．$\lambda (> 0)$ は両項のバランスをとる定数である．

式 (6) は $\boldsymbol{L}$ を用いて，式 (7) と変形できる．$\boldsymbol{L} (\equiv \boldsymbol{D} - \boldsymbol{W})$ はラプラシアン行列と呼ばれ，対角行列 $\boldsymbol{D}$ は $\boldsymbol{W}$ の各行 (又は列) の和を対角成分に持つ行列である．

$$\begin{aligned} J(\boldsymbol{f}) &= \sum_{i=1}^{l} (y^{(i)} - f^{(i)})^2 \\ &\quad +\lambda \sum_{i<j} w^{(i,j)} (f^{(i)} - f^{(j)})^2 \end{aligned} \quad (6)$$

$$= ||\boldsymbol{y} - \boldsymbol{f}||_2^2 + \lambda \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} \quad (7)$$

$$\boldsymbol{f} = (\boldsymbol{I} + \lambda \boldsymbol{L})^{-1} \boldsymbol{y} \quad (8)$$

3 実験

3.1 実験仕様

テキスト分類問題の対象データには，Reuters-21578 (Reuters)[1] を用いる．Reuters は 135 のトピックカテゴリからなる Reuters newswire の英文記事を集めたデータセットである．本実験では “ModApte”分割に従って，本文とタイトルのみからなる記事データを抽出し，全データに対してストップワードの除去とステミング処理を行う．その後，同じデータセットを用いて GBSSL 手法でマルチラベル文書分類を行っている Subramanya ら [4] の実験仕様に合わせ，10 種のカテゴリ **earn**，**acq**，**money-fx**，**grain**，**crude**，**trade**，**interest**，**ship**，**wheat**，**corn** に対する分類精度を求める．Reuters の記事データはマルチラベルを有するため，ここでは各カテゴリ毎に one-versus-rest 法を適用した二値分類を行い，一定の閾値以上のカテゴリラベルを文書に付与するラベルとして採用する．

データセットは，テストデータ (ラベルなしデータ)$u = 3299$ 個を共通とし，これに教師データ $l = 20$ 個を加えたものを 16 セット用意する．データセットに含まれるデータ総数は $n = 3319$ 個である．教師データとして加えるカテゴリは，上記 10 種のカテゴリにそれら以外のカテゴリ (**others**) を加えた全 11 種とする．データ

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

セットに加える教師データ l 個のカテゴリは 11 種のカテゴリからランダムに選択するが，全 11 種のカテゴリの教師データが少なくとも 1 個ずつ含まれるように選択する．

グラフ構成の際に求める，潜在トピックの推定方法には，崩壊型ギブスサンプリングを用い，その反復回数は 200 回とする．最適トピック数はパープレキシティの値を算出し，その 5 回平均の値で決定する．$\alpha = 0$ のときは文書の表層情報のみを扱うため，推定を行う必要がない．このため，類似度が一意的に決まる．他方，$\alpha \neq 0$ のときは文書の潜在トピックの推定を行うため，類似度が一意的に決まらない．このため，5 回平均した値を用いることとする．ノード間の類似度におけるパラメータ α は $[0, 1]$ の範囲を 0.1 刻みで動かす．

ラベル伝搬法で用いた類似度グラフのノード数は $|V| = n(= 3319)$ である．k-近傍グラフの大きさのパラメータ k は $\{2, 10, 50, 100, 250, 500, 1000, 2000, n\}$，ラベル伝搬法のパラメータ λ は $\{1, 0.1, 0.01, 1e{-}4, 1e{-}8\}$ の範囲を動かす．15 セット中 5 つのデータセットによって，各カテゴリに対する最適パラメータ (k, λ) の組を決定した後，それらのパラメータの値を用いて，残り 10 セットに対して文書分類を行い，各カテゴリ毎に PRBEP を求め，各試行毎の各カテゴリに対する PRBEP の平均値を算出する．指標 PRBEP は，$Precision$(適合率) と $Recall$(再現率) が一致するときの値である．

3.2　実験結果

$[0, 1]$ における 0.1 刻み毎の各 α の値に対して，カテゴリ毎に決定した最適パラメータ (k, λ) を表 1 に示す．各カテゴリに対し，これらの最適パラメータを用いて行った実験結果を図 1〜10 に示す．横軸は α の値を表し，縦軸は PRBEP の値を表す．図 1〜10 は，各 α の値に対して行った 10 回の試行の各カテゴリ PRBEP の平均値を示している．各 α 毎に全カテゴリの PRBEP を合算して求め，その平均値の変移を図 11 に示す．図 12，13 は $\alpha = 0, 0.2, 1$ のときの，カテゴリ毎のテストデータ数とその PRBEP との相関関係を表している．横軸は各カテゴリに含まれるテストデータの数を表し，縦軸は PRBEP の値を表す．青の点線，黒の実線そして赤の一点鎖線は，それぞれ $\alpha = 0, 0.2, 1$ のときの結果を表す．

図 1〜11 において，$\alpha = 0$ の場合は，表層情報のみを用いた場合の結果であり，本研究におけるベースラインである．また，$\alpha = 1$ の場合は，潜在情報のみを用いた場合の結果である．それ以外 ($\alpha \neq 0$ または 1) は，潜在情報と表層情報を一定の割合 $(\alpha : (1 - \alpha))$ で混合した場合であり，両情報を用いた結果を示している．

まず，図 1〜10 に関連しては次の通りである．$\alpha = 0$ の時よりも，$\alpha \neq 0$ の時の PRBEP が必ず大きい値をとるのは，図 1, 2, 3, 6, 7, 8 である．他方，逆に $\alpha = 0$ の時よりも，$\alpha \neq 0$ の時の PRBEP が α の値によって小さい値をとるのは，図 4, 5, 9, 10 である．

次に，図 11 からは以下のことが分かる．マクロ平均値の最大値は 51.0 ($\alpha = 0.2$) であり，最小値は 44.5 ($\alpha = 1$) である．ただし，$\alpha = 0$ の時の値は 45.2 である．したがって，最大マクロ平均値 51.0 ($\alpha = 0.2$) は $\alpha = 1$ の時より 6.5%高く，更にベースラインである $\alpha = 0$ の時より 5.9%高いことが分かる．また，$\alpha = 0$ 〜0.2 の時，マクロ平均値は単調増加しており (45.2 → 51.0)，$\alpha = 0.2$ 以上では，マクロ平均値は単調減少している (51.0 → 44.5)．

図 12, 13 は，$\alpha = 0, 1$，並びにマクロ平均値で最大値をとる $\alpha = 0.2$ における，各カテゴリのテストデータ数とその精度の相関関係を表している．テストデータ数の多いカテゴリほど，潜在トピックを考慮した $\alpha \neq 0$ における精度は改善されていることが分かる．しかしながら，データ数が 200 個以下であるカテゴリにおいては必ずしも同様の改善傾向は見られない．

4　考察

図 1〜10 の各図において，PRBEP が最大値をとる時の α の値は各カテゴリ毎に異なっており，一律ではない．故に，精度が最大となる時の，α の値 (すなわち表層情報と潜在情報の混合割合) を一意的に決めることは難しい．しかしながら，半数以上のカテゴリにおいては，$\alpha = 0$ に対して $\alpha \neq 0$ のときの PRBEP は増加傾向を示しており，残りのカテゴリにおいても，適切な α が求まりさえすれば全てのカテゴリにおいてベースラインを超えることが分かる．

図 11 は，全カテゴリのマクロ平均 PRBEP を示している．$\alpha = 1$ を除いた全ての $\alpha \neq 0$ において，ベースラインである $\alpha = 0$ の時のマクロ平均値よりも高くなっている．特に，各 α をベースラインと比較すると，$\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$ のとき，t 検定によって 5%有意でベースラインに対して精度向上があることが分かった．

図 12, 13 からは，ノード間の類似度に文書間の潜在トピックを考慮することが GBSSL の精度改善に繋がることが期待され、それは特に各カテゴリのデータ数が十分多量にあるときであるということが期待される．カテゴリ **wheat**, **corn** において，$\alpha = 0$ に対して $\alpha = 1$ のときの PRBEP が著しく悪化したのは，これらのカテゴリにおけるテストデータが少量であるため，LDA による十分なトピック推定が行えなかったためだと考えられる．

以上のことから，GBSSL 法のグラフ構成としては，表層情報のみを用いるよりも潜在情報も加えた両情報を

表 1: カテゴリ毎の最適パラメータ (k, λ)

カテゴリ \ α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
earn	(50, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)
acq	(500, 0.1)	(250, 0.1)	(250, 0.01)	(100, 0.01)	(100, 1e-8)	(50, 0.1)	(10, 1e-8)	(10, 1e-8)	(10, 1e-4)	(250, 0.01)	(500, 1e-4)
money-fx	(2, 1)	(2, 1)	(10, 0.1)	(2, 0.1)	(2, 1)	(2, 1)	(50, 1e-4)	(50, 0.01)	(2, 1e-8)	(50, 0.01)	(10, 0.1)
grain	(100, 0.1)	(50, 1)	(50, 1)	(10, 1)	(50, 1e-8)	(10, 1)	(10, 1)	(50, 1e-8)	(50, 1e-8)	(50, 1)	(50, 1)
crude	(10, 1)	(50, 0.1)	(50, 0.01)	(100, 1e-8)	(10, 0.01)	(10, 1e-8)	(50, 1e-8)	(2, 1e-4)	(50, 1e-8)	(2, 1e-8)	(50, 1e-8)
trade	(10, 1)	(10, 1e-8)	(10, 1e-8)	(10, 1e-4)	(10, 1e-8)	(10, 1e-4)	(10, 1e-8)	(2, 0.01)	(10, 1e-8)	(10, 1e-8)	(10, 0.1)
interest	(10, 0.1)	(10, 1)	(10, 0.1)	(10, 1e-8)	(10, 1)	(10, 1)	(10, 1)	(10, 1)	(10, 1)	(100, 1e-8)	(100, 1e-8)
ship	(10, 1)	(100, 1e-8)	(50, 0.1)	(10, 1e-8)	(10, 0.1)	(10, 0.1)	(10, 0.1)	(10, 0.1)	(2, 1)	(10, 0.1)	(10, 0.1)
wheat	(100, 0.01)	(100, 1e-8)	(100, 1e-8)	(50, 1e-4)	(50, 1e-4)	(50, 1e-4)	(100, 1e-8)	(50, 1e-8)	(50, 1e-8)	(50, 1e-8)	(50, 1e-8)
corn	(10, 1)	(10, 1)	(10, 1)	(10, 1)	(10, 1)	(10, 0.01)	(10, 0.01)	(10, 0.1)	(10, 0.1)	(2, 1e-8)	(10, 1e-8)

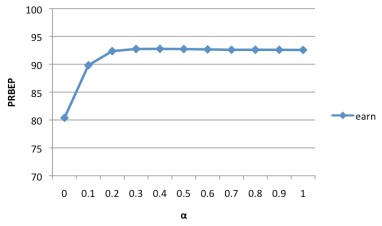

図 1: **earn** の平均 PRBEP

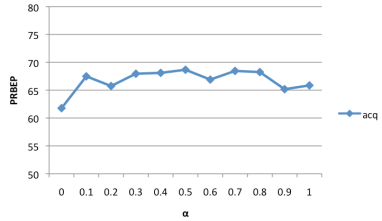

図 2: **acq** の平均 PRBEP

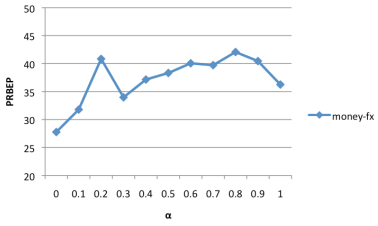

図 3: **money-fx** の平均 PRBEP

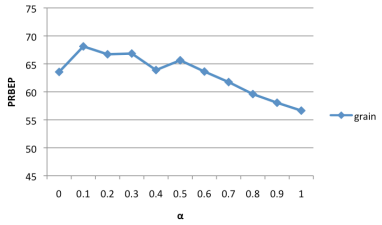

図 4: **grain** の平均 PRBEP

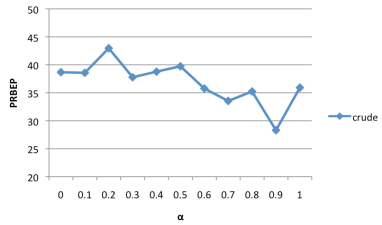

図 5: **crude** の平均 PRBEP

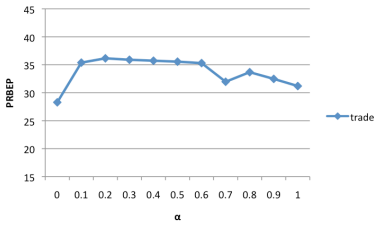

図 6: **trade** の平均 PRBEP

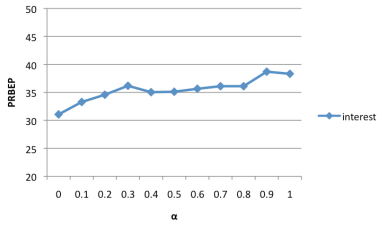

図 7: **interest** の平均 PRBEP

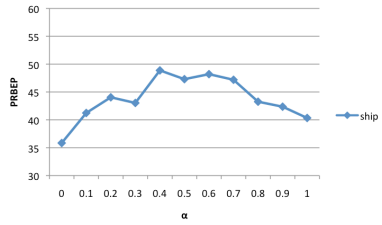

図 8: **ship** の平均 PRBEP

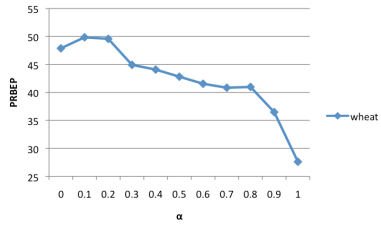

図 9: **wheat** の平均 PRBEP

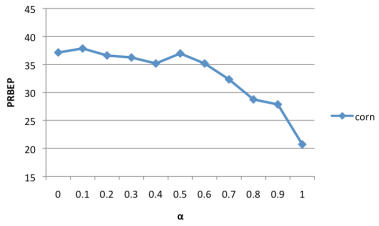

図 10: **corn** の平均 PRBEP

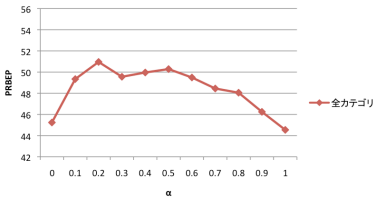

図 11: 全カテゴリのマクロ平均 PRBEP

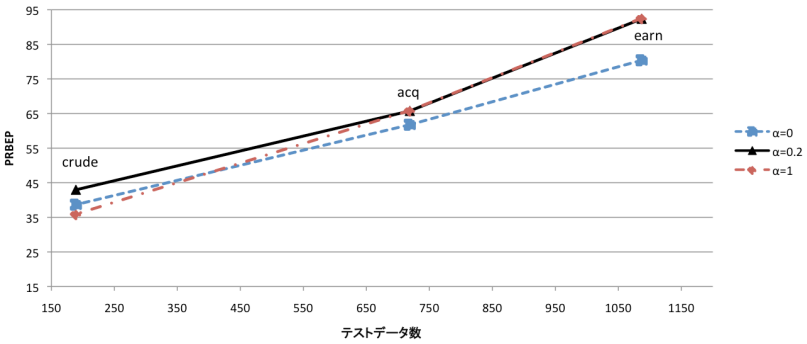

図 12: 各カテゴリにおけるテストデータ数と PRBEP との相関関係．$\alpha = 0, 0.2, 1$ における，カテゴリ **earn**，**acq**，**money-fx** のテストデータ数 (横軸) と PRBEP(縦軸)

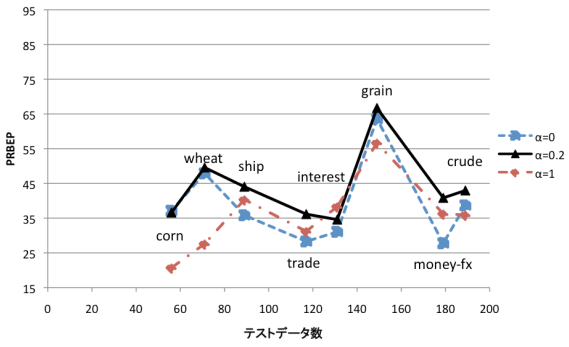

図 13: 各カテゴリにおけるテストデータ数と PRBEP との相関関係．$\alpha = 0, 0.2, 1$ における，カテゴリ **money-fx**，**grain**，**crude**，**trade**，**interest**，**ship**，**wheat**，**corn** のテストデータ数 (横軸) と PRBEP(縦軸)

用いる方がGBSSL法の精度は向上することが分かる．また，十分なデータ数があるときのみ，潜在情報による精度向上への寄与率が上がることも期待される．したがって，両情報の混合割合αの最適値が求まり，各カテゴリそれぞれにおいて十分な量のテストデータがありさえすれば，単に表層情報や潜在情報のみを用いる場合よりも，高い精度が得られるだろう．

5　結論

我々は，表層情報と潜在情報に基づく類似度グラフの構成法を提案した．マルチラベルを有するReuters-21578コーパスを用いた実験の結果から，GBSSL法におけるグラフ構成では表層情報と潜在情報のどちらかだけを用いるよりも，両情報を混合させて同時に用いた方がGBSSL法における文書分類の精度を向上させることが分かった．

今後の課題としては，我々が今回得た結論(表層情報と潜在情報の両情報を用いる方がそれらを単体で用いるよりも精度が高い)を他のデータセットを用いて検証することであり，グラフスパース化手法などの工夫を行うことなどを通して更なる精度の向上を図ることである．

参考文献

[1] Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research* (2003)

[2] Cortes, C., Vapnik, V.: Support-vector networks, *Machine Learning*,20: 273-297 (1995)

[3] Salton, G., McGill, J.: Introduction to Modern Information Retrieval, McGraw-Hill (1983)

[4] Subramanya, A., Bilmes, J.: Soft-Supervised Learning for Text Classification, in *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp.1090–1099 (2008)

[5] Zhou, D., Bousquet, O., Lal, T. N., Weston J., Schölkopf B.: Learning with Local and Global Consistency, in *NIPS 16* (2004)

[6] Zhu, X., Ghahramani, Z.: Learning from Labeled and Unlabeled Data with Label Propagation, Technical report, Carnegie Mellon University (2002)

[7] Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions, in *Proc. of the International Conference on Machine Learning (ICML)* (2003)

[8] Zhu, X., Ghahramani, Z., Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions, In *ICML* (2003)

[9] Zhu, X.: Semi-Supervised Learning with Graphs, PhD thesis, Carnegie Mellon University (2005)

[10] Gu, Q. and Han, J.: Towards Active Learning on Graphs: An Error Bound Minimization Approach, Data Mining, IEEE International Conference (2012)

[11] Ozaki, K., Shimbo, M., Komachi, M. and Matsumoto, Y.: Using the mutual k-nearest neighbor graphs for semi-supervised classification of natural language Data, Proceedings of the Fifteenth Conference on Computational Natural Language Learning (2011)

[12] Jebara, T., Wang, J. and Chang, S.: Graph construction and b-matching for semi-supervised learning, Proceedings of the 26th Annual International Conference on Machine Learning (2009)