潜在情報を考慮したグラフに基づく教師データの選出による ラベル伝搬法

江里口 瑛子　　小林 一郎
お茶の水女子大学 理学部 情報科学科
{g0920506, koba}@is.ocha.ac.jp

1　序論

機械学習手法には教師あり学習，教師なし学習，半教師あり学習等があり，その中でグラフ構造に基づく半教師あり学習 (Graph-based Semi-Supervised Learning: GBSSL) 法は，SVM 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 [1]．GBSSL 法の精度は，一方で，グラフ構成の仕方によって左右される．他方，その精度はどのようなラベルありデータ (教師データ) を与えるかによっても左右される．特に，後者のデータの選出に関連しては，能動学習法を用いた質の高いラベルありデータを選出する手法が，少ない数のラベルありデータを用いて，GBSSL 法の精度を向上させることが知られている [2]．

本研究では，後者のラベルありデータの選出によって精度を向上させるため，グラフの構成に工夫を凝らし，それを用いて質の高いラベルありデータを選出する．一般に，テキストデータからなるグラフを構成する際には，これまで文書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，文書間の潜在的な類似度を加えたものを採用する．この潜在情報を考慮したグラフ上で，更に PageRank[3] 手法を用いて，質の高いラベルありデータを選出できるようにする．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，PRBEP を算出し，こ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　グラフに基づく文書分類手法

2.1　グラフ構成

本研究では，ノード間の類似度を重みとする重み付き無向グラフ $G = (V, E)$ を用いる．ここで 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

一般に，テキストデータを対象にしたグラフ構成においては，ノード間の類似度に，文書の表層情報に基づく類似度として，単語の出現頻度に着目した *tfidf* の *cos* 類似度が多く用いられる．本研究では，この従来の類似度に新たに，文書の持つ潜在情報に基づいた類似度を　(0　　1) の割合で付加し，これらを合算してノード間の類似度とする (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．文書内のトピック分布推定手法には，Latent Dirichlet Allocation(LDA)[4] を用い，トピック分布の類似度指標には，式 (2) によって Jensen-Shannon ダイバージェンス (D_{JS}) を類似度に変換したものを用いる．

$$sim(S,T) \equiv \quad * sim(P,Q) + \quad (1 - \quad) * cos(tfidf(S), tfidf(T)) \tag{1}$$

$$sim(P,Q) \equiv 1 - D_{JS}(P,Q) \tag{2}$$

2.3　ラベルありデータの選出

ラベルありデータの選出は，北島ら [5] によって提案された TopicRank 法を採用して行う．TopicRank 法とは，ノードを文とし，文間の潜在情報に基づく類似度で構成されたグラフに対して，式 (3) を用いて文の重要度を算出し，順位付けを行う手法である．ここで，d は制動係数 (damping factor) である．

本研究では，グラフのノードを文から文書 (文の集合) に置き換えるため，式 (3) において，N を対象文書

群の総文書数，$adj[u]$ を文書 u の隣接ノード集合とする．その上で，文書のトピック分布を考慮した，ラベルありデータのみをノードにもつグラフをカテゴリ毎に作成し，TopicRank スコアが高いデータから順に，GBSSL 法で用いるラベルありデータとしていく．

$$r(u) = d \sum_{v \in adj[u]} \frac{sim(u,v)}{\sum_{z \in adj[v]} sim(z,v)} p(u) + \frac{1-d}{N} \quad (3)$$

2.4 ラベル伝搬法

本研究における学習法として，ラベル伝搬法 [6, 7] を採用する．「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について予測を行う．類似度行列を $\boldsymbol{W}$，ノード数を n 個 (このうちラベルありデータ数は 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 (4)) の解 (式 (5)) として求まる．$\boldsymbol{L}(\equiv \boldsymbol{D} - \boldsymbol{W})$ 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．

$$\begin{aligned} J(\boldsymbol{f}) &= \sum_{i=1}^{l} (y^{(i)} - f^{(i)})^2 + \sum_{i<j} w^{(i,j)} (f^{(i)} - f^{(j)})^2 \\ &= ||\boldsymbol{y} - \boldsymbol{f}||_2^2 + \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} \end{aligned} \quad (4)$$

$$\boldsymbol{f} \equiv (\boldsymbol{I} + \boldsymbol{L})^{-1} \boldsymbol{y} \quad (5)$$

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)[8] 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 [1] 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，これにラベルありデータ $l = 20$ 個を加えたものを 11 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．ラベルあり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．データセットに加えるラベルありデータ 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 種のカテゴリのラベルありデータが少なくとも 1 個ずつ含まれるように選択する．

TopicRank 法を用いる際の LDA 法における潜在トピックの推定方法には，ギブスサンプリングを用い，その反復回数は 200 回とする．トピック数はパープレキシティの値を算出し，その 10 回平均の値で決定する．また，TopicRank 法で用いるグラフは，ノード数 $|V|$ =(カテゴリ毎のラベルありデータの総数)，辺数 $E = |V \times V|$ の完全グラフとする．パラメータ　は，0.0 から 1.0 まで 0.1 刻み毎の値を与え，制動係数 d は Brin ら [3] の結果を参考に 0.85 とする．カテゴリ毎に各文書の TopicRank スコアを算出し，テストデータに加えるラベルありデータのカテゴリ数にしたがって，スコアの高いラベルありデータから順にデータセットに加えていく．　$= 0$ のときは文書の表層情報のみを扱い，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のでスコアは 1 回のみ算出する．他方，　$\neq 0$ 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 回平均の値をスコアとする．

ラベル伝搬法で用いた類似度グラフのノード数は $|V| = n(= 3319)$ であり，ノード間の類似度は，パラメータ　$= 0$ とし，表層情報のみからなるものとする．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　は $\{1, 0.1, 0.01, 1e-4, 1e-8\}$ の範囲を動かす．最初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 $(k,\ )$ 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　の値に対して，カテゴリ毎に決定した最適パラメータ $(k,\ )$ を表 1 に示す．

表 1: カテゴリ毎の最適パラメータ (k, α)

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	(500, 1)	(50, 1)	(1000, 1)	(1000, 1)	(50, 1)	(50, 1)	(50, 1)	(50, 1)	(50, 1)	(50, 1)	(50, 1)
acq	(100, 0.01)	(100, 0.01)	(100, 0.01)	(2, 1)	(100, 0.01)	(100, 0.01)	(100, 1e-8)	(100, 1e-8)	(100, 1e-8)	(100, 1e-8)	(100, 1e-8)
money-fx	(250, 0.01)	(100, 1e-8)	(10, 1e-4)	(100,1e-8)	(2, 0.1)	(2, 0.1)	(2, 1e-8)	(250, 1e-8)	(2, 0.1)	(2, 1e-8)	(250, 1e-8)
grain	(250, 0.1)	(2000, 1e-4)	(100, 1)	(250, 0.1)	(100, 1)	(50, 1)	(250, 1)	(50, 1)	(50, 1)	(50, 1)	(100, 1)
crude	(50, 0.1)	(2, 1)	(200, 0.01)	(50, 1e-8)	(10, 0.01)	(250, 0.01)	(250, 0.01)	(250, 1e-8)	(10, 0.01)	(250, 0.01)	(250, 0.01)
trade	(2, 1)	(10, 0.1)	(50, 0.01)	(10, 1e-8)	(10, 1e-8)	(10, 1e-8)	(50, 1e-8)	(10, 1e-8)	(10, 1e-4)	(10, 0.1)	(10, 0.1)
interest	(10, 1)	(50, 1e-8)	(50, 1e-8)	(10, 1)	(2, 0.1)	(250, 1e-8)	(250, 0.01)	(250, 0.01)	(2,1)	(2, 0.1)	(500, 1e-8)
ship	(3318, 1)	(50, 1)	(50, 1)	(250, 0.1)	(50, 0.1)	(50, 0.1)	(50, 1e-8)	(50, 1e-8)	(100, 0.1)	(100, 0.1)	(50, 0.01)
wheat	(500, 1e-8)	(500, 1e-8)	(250, 1e-8)	(500, 1e-8)	(500, 0.01)	(1000, 0.01)	(500, 1e-8)	(250,1e-8)	(250, 1e-8)	(250, 1e-8)	(250, 1e-8)
corn	(10, 1e-8)	(100, 1e-8)	(250, 1e-8)	(10, 1e-8)	(250, 1e-8)	(250, 1e-4)	(500, 1e-8)	(100, 1e-8)	(250, 1e-8)	(50, 0.01)	(250, 1e-4)

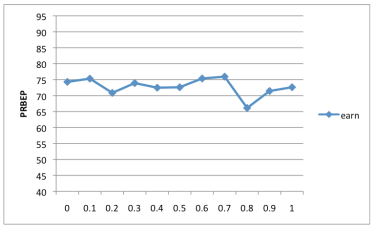

図 1: **earn** の平均 PRBEP

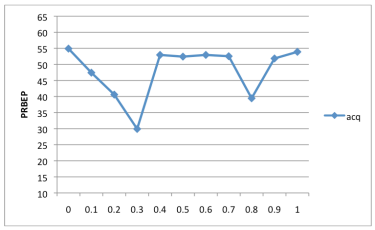

図 2: **acq** の平均 PRBEP

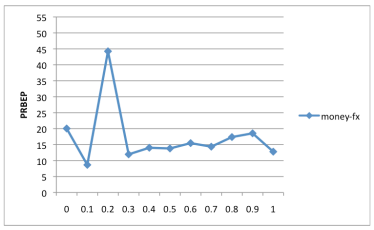

図 3: **money-fx** の平均 PRBEP

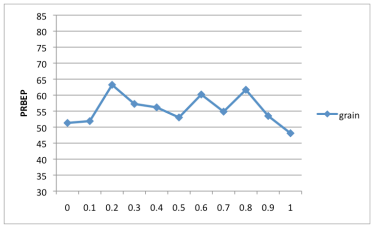

図 4: **grain** の平均 PRBEP

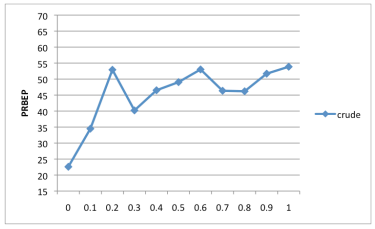

図 5: **crude** の平均 PRBEP

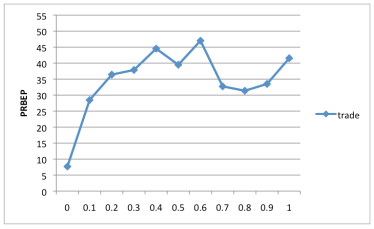

図 6: **trade** の平均 PRBEP

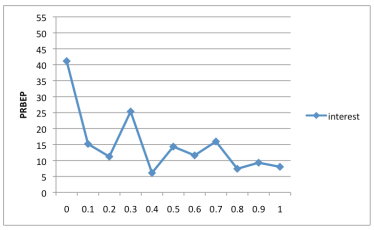

図 7: **interest** の平均 PRBEP

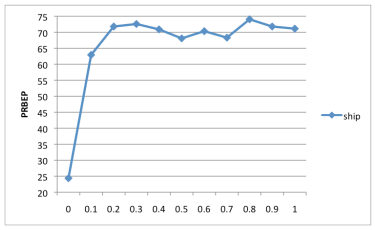

図 8: **ship** の平均 PRBEP

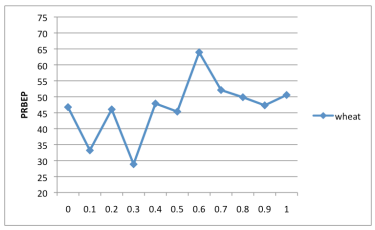

図 9: **wheat** の平均 PRBEP

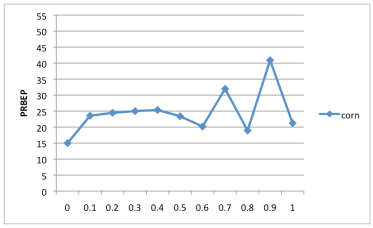

図 10: **corn** の平均 PRBEP

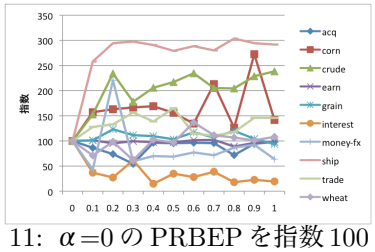

図 11: α=0 の PRBEP を指数 100 とした時の PRBEP の割合

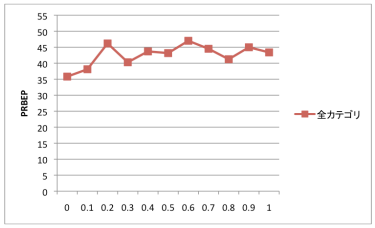

図 12: 全カテゴリの平均 PRBEP

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．図 1〜10 は各 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．図 11 は　= 0 の PRBEP を指数 100 とした際の，各　に対する各カテゴリにおける PRBEP の割合の変移を示している．各　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 12 に示す．各図の横軸は全て　の値を表す．

　= 0 の場合は，表層情報のみを用いた場合の結果であり，それ以外 (　≠ 0) は，潜在情報と表層情報を一定の割合 (　: (1 −　)) 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．

図 1〜10 に関連しては次の通りである．　が増加するにつれて PRBEP に増加傾向が見られるのは，図 4，5，6，8，10 である．他方，逆に　が増加するにつれて PRBEP が減少しているものは，図 2，7 である．　= 0 の時の値に対して，PRBEP が上下の変動を繰り返し，一意的な相関関係を見て取るのが難しいものは，図 1，3，9 である．

図 11 から分かるように，　が増加するにつれて，正の相関が見られるものに関連して，PRBEP が最善で正方向に 200% も増加し，精度の向上が見られるものもあれば，他方，負の相関が見られるものもあり，最悪で約 80% 減殺され，PRBEP が減少しているものもある．図 12 からは，各カテゴリにおける PRBEP を総計してその平均値の変移を見ると，　が増加するにつれ，PBREP が増加する傾向があることが見て取れる．

3.3　考察

　= 0 に対して，　= 0.1 以上でカテゴリ毎の PRBEP の向上が見られたものは全カテゴリの半数あり，かつ，全体で見たときも精度の向上が見られ，これらの点において我々の手法がある一定程度有効であると言えよう．しかし，期待に反した PRBEP の挙動を示すものもあり，問題を残した．これに関連しては，実験手法の都合により，カテゴリ毎の最適パラメータ (k,　) を 1 つのデータセットだけを用いて決定したことや，潜在情報の活用をラベルありデータの選出時のみに留め，ラベル伝搬時には用いず，選出されたデータの表層情報のみに限定したことなどが要因として考えられる．

4　結論

GBSSL 法において，ラベルありデータの選出法として，潜在情報を考慮したグラフを構成し，このグラフに対して PageRank アルゴリズムを用いて，その後にマルチラベル分類を行う我々の手法は，限定付きであるが，精度の向上が見られ，マルチラベルを有するデータに対する分類において，ある一定程度有効である．

今後の課題としては，最適パラメータ (k,　) における決定の仕方を改善することや，ラベル伝搬法で用いるグラフ構成にも潜在情報を活用することによって，更なる精度の向上を図ることである．

参考文献

[1] A. Subramanya and J. Bilmes, Soft-supervised text classi cation, In *EMNLP*. 2008.

[2] X. Zhu, J. La erty, and Z. Ghahramani, Combining active learning and semi-supervised learning using Gaussian elds and harmonic functions, In *ICML* workshop, 2003.

[3] S. Brin and L. Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems, 30, pp. 107-117, 1998.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022, 2003.

[5] 北島 理沙, 小林 一郎, 潜在的意味を考慮したグラフに基づく複数文書要約, Proceeding of ARG WI2, 2012.

[6] D. Zhou, O. Bousquet, J. Weston, and B. Schölkopf, Learning with local and global consistency, In *NIPS 16*, 2004.

[7] X. Zhu, Z. Ghahramani, and J. La erty, Semi-supervised learning using Gaussian elds and harmonic functions, In *ICML*, 2003.

[8] D. Lewis, Reuters-21578 text categorization test collection distribution 1.0, `http://www.daviddlewis.com/resources/testcollections/reuters21578/`, 1997.