半教師あり学習における教師データ選出とグラフ構成

High-Quality Training Data Selection and Graph Construction for Graph-based Semi-supervised Learning

江里口 瑛子 [1] *　　小林 一郎 [2]
Akiko Eriguchi[1]　　Ichiro Kobayashi[2]

[1] お茶の水女子大学理学部情報科学科
[1] Department of Information Sciences, Ochanomizu University
[2] 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
[2] Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Abstract: We try raising the accuracy of multi-class document categorization using graph-based semi-supervised learning (GBSSL). With this end in view, we propose two methods. The first one is a method to construct a similarity graph by employing both surface information and latent information to express similarity between nodes. The second one is a method to select high-quality training data for GBSSL by means of PageRank algorithm. We experimented on Reuters-21578 corpus. We have confirmed that our proposed methods work well for raising the accuracy of multi-class document categorization.

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．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，Support Vector Machine (SVM)[4] 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 [6]．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，他方で，その精度は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も左右される．前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 [11]．後者に関連して重要となるのが，情報量の大きい教師データをどのように選出するかである．その良い事例が能動学習法であり，これは教師データの数は少ないが，質の高い教師データを選出する方であり，これによって GBSSL 法の精度が向上することが知られている [9]．

本研究は，多クラス文書分類における GBSSL 法の精度向上を目指すものであり，この手段として二つの方法を提示する．

まず第一は，グラフ構成に関連するものであり，グラフ構成に必須の要件である類似度に，文書間の潜在的な類似度を取り入れる．一般にこれまで，テキストデータからなるグラフを構成する際には，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，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．

第二は，教師データの選出に関連するものであり，教師データからなる類似度グラフにおいて，各ノードにスコアを付けて質の高い教師データを選出する．すなわち，潜在情報を加味したグラフ上で，PageRank[2] 手法を用いて，質の高い教師データを選出する．ちなみに，ここで用いるグラフのノード間の類似度は，先述の文書間の潜在的な類似度を加味したものである．

以上の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，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．

*連絡先： お茶の水女子大学理学部情報科学科
〒 112-8610 東京都文京区大塚 2-1-1
E-mail: g0920506@is.ocha.ac.jp

2　グラフに基づく文書分類手法

2.1　グラフ構成

本研究におけるグラフ構成においては，テキストデータを対象にしたグラフ構成を行う．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 [5] が多く用いられる．ここでは，表層情報に基づく類似度を，*tfidf* ベクトルのコサイン類似度 (sim_{cos}) 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を，式 (2) によって得られる値 (sim_{JS}) とし，トピック分布間の距離は Jensen-Shannon ダイバージェンス (D_{JS}) 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_{cos}) に新たに，文書の持つ潜在情報に基づいた類似度 (sim_{JS}) を $\alpha(0 \leq \alpha \leq 1)$ の割合で付加する．これら sim_{JS} と sim_{cos} 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．

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

$$sim_{JS}(P,Q) \equiv 1 - D_{JS}(P,Q) \quad (2)$$

2.3 質の高い教師データの選出

質の高い教師データの選出法として，北島ら [12] によって提案された TopicRank 法を採用して行う．TopicRank 法とは，グラフ構造を用いた重要文抽出法の一つであり，類似度グラフのノードを単文とし，辺の重みを文間の潜在情報に基づく類似度として構成したグラフに対して，PageRank[2] の概念を用いて式 (3) により各ノード (各単文) の重要度を算出し，各ノードの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，類似度グラフのノードを単文から文書 (文の集合) に置き換えて用いることとする．このため，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．$sim_{nodes}(u,v)$ は，式 (1) によって求めた文書 u と文書 v の類似度である．その上で，文書のトピック分布を考慮した，教師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 法で用いる教師データとしていく．

$$\begin{aligned} r(u) &= d \sum_{v \in adj[u]} \frac{sim_{nodes}(u,v)}{\sum_{z \in adj[v]} sim_{nodes}(z,v)} r(u) \\ &\quad + \frac{1-d}{N} \quad (3) \end{aligned}$$

2.4 ラベル伝搬法

本研究における GBSSL 法として，ラベル伝搬法 [7, 10] 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 (4)) の解 (式 (6)) として求まる．式 (4) 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 (4) は $\boldsymbol{L}$ を用いて，式 (5) 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．

$$\begin{aligned} J(\boldsymbol{f}) &= \sum_{i=1}^{l} (y^{(i)} - f^{(i)})^2 \\ &\quad + \lambda \sum_{i<j} w^{(i,j)} (f^{(i)} - f^{(j)})^2 \quad (4) \\ &= ||\boldsymbol{y} - \boldsymbol{f}||_2^2 + \lambda \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} \quad (5) \end{aligned}$$

$$\boldsymbol{f} = (\boldsymbol{I} + \lambda \boldsymbol{L})^{-1} \boldsymbol{y} \quad (6)$$

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 [6] 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 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．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．

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 $|V|$ =(カテゴリ毎の教師データの総数)，辺数 $E = |V \times V|$ の完全グラフとする．パラメータ α 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 [2] 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教師データのカテゴリ数にしたがって，スコアの高い教師データから順にデータセットに加えていく．$\alpha = 0$ 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，$\alpha \neq 0$ 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 $\alpha = 0$ 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 λ 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, 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．

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

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．図 11 は $\alpha = 0$ の PRBEP を指数 100 とした際の，各 α 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 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 12 に示す．図中のエラーバーは標準偏差を表す．

全ての図において，$\alpha = 0$ 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 が大きい値をとるのは，図 4，5，6，8，10 である．ただし，図 4 では，$\alpha = 1$ の時の PRBEP は $\alpha = 0$ の時よりも小さい値をとる．他方，逆に $\alpha = 0$ の時よりも，$\alpha \neq 0$ の時の PRBEP が小さい値をとるのは，図 2，7 である．そして，$\alpha = 0$ の時の値に対して，$\alpha \neq 0$ の時の 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．

次に，図 11 から分かるように，α 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．

最後に，図 12 からは以下のことが分かる．マクロ平均値の極大値は 46.2，46.9，45.0(それぞれ $\alpha = 0.2, 0.6, 0.9$) であり，最大値は 46.9($\alpha = 0.6$ の時) である．最小値は

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	(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)	(250, 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)

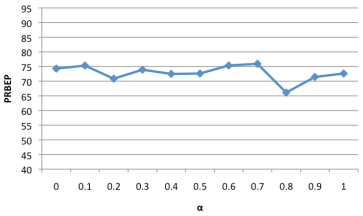

図 1: **earn** の平均 PRBEP

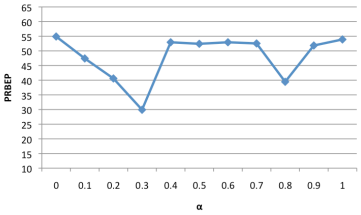

図 2: **acq** の平均 PRBEP

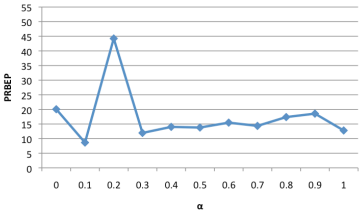

図 3: **money-fx** の平均 PRBEP

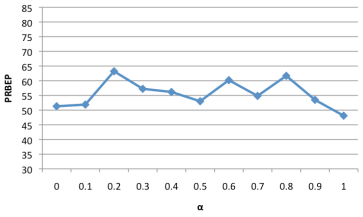

図 4: **grain** の平均 PRBEP

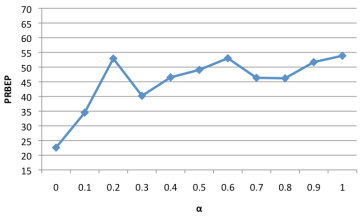

図 5: **crude** の平均 PRBEP

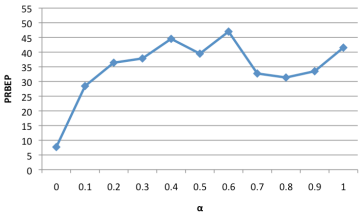

図 6: **trade** の平均 PRBEP

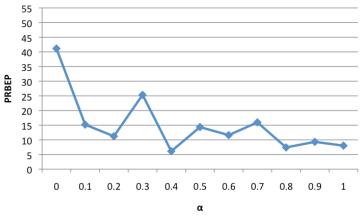

図 7: **interest** の平均 PRBEP

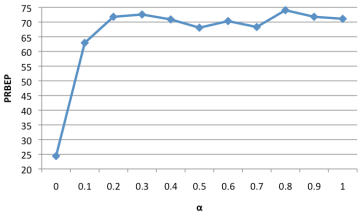

図 8: **ship** の平均 PRBEP

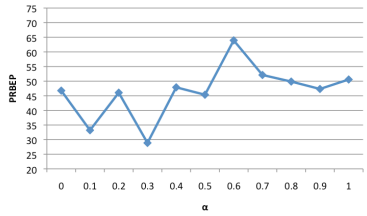

図 9: **wheat** の平均 PRBEP

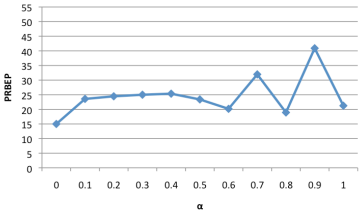

図 10: **corn** の平均 PRBEP

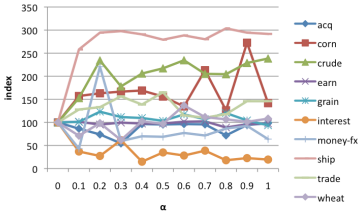

図 11: α=0 の PRBEP を指数 100 とした時の PRBEP の割合

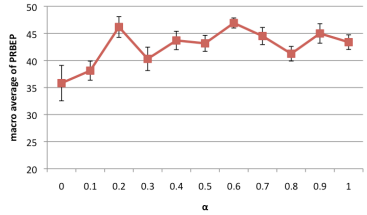

図 12: 全カテゴリの平均 PRBEP

35.8($\alpha = 0$) である．ただし，$\alpha = 1$ の時の値は 43.4 である．したがって，最大マクロ平均値 46.9($\alpha = 0.6$) は $\alpha = 1$ の時より 3.5%高く，更に $\alpha = 0$ の時より 11.1%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 (35.8 → 46.2)，$\alpha = 0.2$ 以上では，マクロ平均値は一定の範囲 (40.3〜46.9) を浮動している．以上から，$0.1 \leq \alpha \leq 1$ の時のマクロ平均値は $\alpha = 0$ の時よりも大きい．更に重要なことは，$\alpha = 0.2, 0.4, 0.6, 0.7, 0.9$ の時のマクロ平均値は，$\alpha = 0$ の時よりも大きいことは勿論，$\alpha = 1$ の時よりも大きいということである．

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，全体的に見た場合には，一定の傾向性や関係性が見て取れる．図 11 においては，半数以上のカテゴリで PRBEP は増加傾向を示している．更に，大半のカテゴリにおいて，$\alpha \geq 0.1$ における PRBEP は $\alpha = 0$ の時の値 100 より大きい．この故に，全カテゴリのマクロ平均値をとった時には右肩上がりの変化パターンが期待される．図 12 は，全カテゴリのマクロ平均 PRBEP が示している．$\alpha = 0$ の時のマクロ平均値をベースラインとすると，$\alpha = 1$ の時の値は 7.6%増加しており，$\alpha = 0.6$ の時の最大値では 11.1%も増加している．加えて，$\alpha = 0.2, 0.4, 0.6, 0.7, 0.9$ の時の値は $\alpha = 1$ の時の値よりも大きい．

以上のことから，教師データの選出を行う際には，表層情報のみを用いるよりも潜在情報を用いる方が GBSSL 法の精度は向上することが分かる．また，両情報を用いる方が，潜在情報のみを用いるよりも精度は向上する．したがって，両情報の混合割合 α の最適値が求まりさえすれば，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，並びに，GBSSL 法で用いる質の高い教師データの選出法を提案した．Reuters-21578 コーパスを用いた実験の結果から，教師データの選出には，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，最適パラメータ (k, λ) における決定の仕方を改善することであり，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] Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research* (2003)

[2] Brin, S., Page, L.: The Anatomy of a Large-scale Hypertextual Web Search Engine., *Computer Networks and ISDN Systems*, pp. 107–117 (1998)

[3] Erkan, G., Radev, D. R.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, *Journal of Artificial Intelligence Research 22*, pp.457-479 (2004)

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

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

[6] 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)

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

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

[9] 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)

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

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

[12] 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)