Web ページの流動性に即した Web ディレクトリの自動改変

福本 文代　　山梨大学大学院総合研究部　教授
鈴木 良弥　　山梨大学大学院総合研究部　教授

1 はじめに

情報通信技術の発展により，インターネットが日常的に利用されている．Google に代表されるインターネット検索では，分野に関するディレクトリを利用した検索が用いられている．これは大量の Web 文書をディレクトリの各ノード(分野)に予め分類しておくことで，検索性能と精度の向上を目指す手法である．しかし，インターネット上では日々新しい情報が配信されるため，時間が経つにつれて既存の分類体系に合わない文書が現れる．この問題に対処するためには，人手によりディレクトリを修正する必要があるが多大なコストと労力を要する．本研究は，既存の分類体系に合致しない文書について，新たな分野を自動的に推定する手法を提案する．具体的には，Web ページ文書を分類するために，(1) 時間差適応を行う分類モデルを提案する．さらに，(2) 2 つのディレクトリを対象とし，双方におけるディレクトリの整合性を保持しつつ，これらを統合することにより，新たな分野を推定する手法を提案する．

2 時間差適応のための分類モデルの構築

訓練データとテストデータの作成年が異なる文書に対し，転移学習を導入することにより高精度な分類を目指す．転移学習による文書分類の手法の一つに Dai らが提案した TrAdaBoost がある[Dai'16]．しかし，TrAdaBoost は訓練文書とテスト文書それぞれの作成年差を考慮した手法とはなっていない．Dai らは，作成年差の大小に関わらずに誤って分類された訓練データの重みを一律に下げてしまうためである．すなわち作成年差に関わらない訓練データの重みの下げ幅は，作成年差が 1 年や 2 年など小さい場合，また作成年差が 10 年，あるいは 15 年と大きい場合でも一定である．しかし作成年差が小さい場合，訓練データとテストデータの単語の出現傾向の差は少ないことから，分類精度はそれほど低下することはないと考えられる．したがって，Salles らが提案した TWF のように作成年差に応じた学習をする必要がある[Salles]．本研究では，新たに転移学習に作成年差に対応して減少させる重みが変化する関数 twf(z)を用いた分類手法を提案する．さらに，訓練データと作成年が異なるテストデータの自動分類精度を向上させるために，訓練データとテストデータの作成年差を考慮する Word2Vec を用いた意味的な素性(Word2Vec 素性)を追加する．加えて Word2Vec を用いたテストデータに出現するが訓練データに出現しない単語(未知語と呼ぶ)を訓練データに出現した類似語で置き換えるスムージングを行う．Word2Vec 素性を追加し，未知語に対してスムージングを適用することにより作成した新しいデータセットに，新しい学習手法を適用することにより訓練データと作成年が異なるテストデータの高精度な分類を目指す．本手法の枠組みを図 1 に示す。図 1 において *Td* は，テストデータとは年度が異なるデータ(diff 訓練データ)を示し，*Ts* はテストデータと同年であるデータ(same 訓練データ)を示す．

2.1 単語のスムージング

本研究では，Word2Vec を用い，テストデータに出現する単語のスムージング を行う．これにより“グロス”と”メッセンジャー”や”ブッシュ”と”オバマ”などの人名やその時期に起きた出来事の単語出現傾向の変化に対応する．対象とする単語はテストデータにのみ出現し，訓練データには出現しない未知語である．未知語は訓練データに出現しないため，学習時に全く考慮されない素性となってしまう．未知語の数は作成年差が大きくなるほど多くなり，分類精度の低下の一因となっていると考えられる．そこで，この未知語のスムージングを行うことにより文書分類の精度向上を目指す．Word2Vec による素性追加を図 2 に示す．

図 2 において，先ず Word2Vec を用いモデルを学習する．得られた学習モデルを用い，未知語のスムージングを行う．テストデータ S に出現した各未知語に対し，diff 訓練データ Td，あるいは same 訓練データ *Ts* に出現しない単語をスムージング対象とする未知語(Wtarget)とする．訓練データに出現する全ての単語を Wtrain とすると，スムージング対象となる単語 *wi* は，$wi \in$ Wtarget，かつ $wi \notin$ Wtrain である．Word2Vec

を用い，W の J 次元ベクトル表現を獲得する．次にこの単語のベクトル表現を用い．Wtarget と余弦尺度による類似度の値が高い単語を検索する．スムージングする単語を wi とし $wi \in$ Wtarget，かつ $wi \notin$ Wtrain であるとき，wi を $wt \in$ Wtrain に置き換える．このとき，あらかじめ設定した閾値よりも高く，かつ最も余弦尺度の大きい単語 wt を選択し未知語と置き換える．

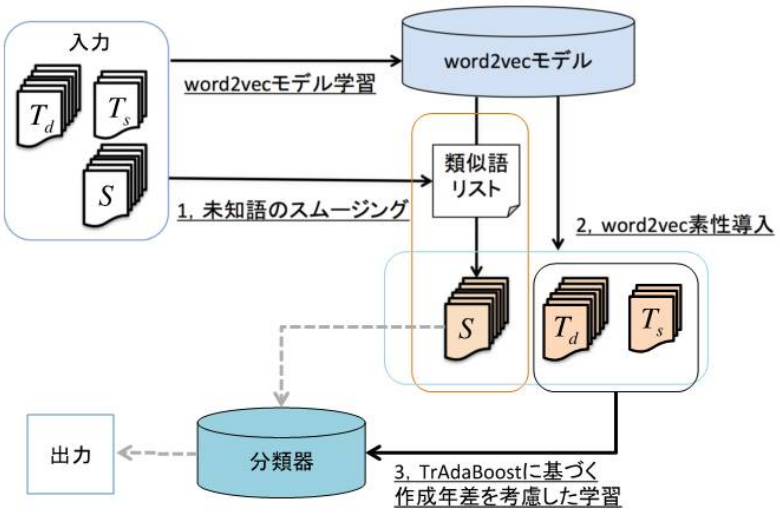


図１　提案手法の流れ

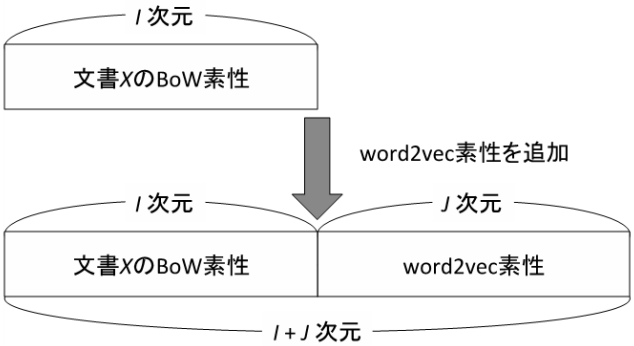


図２　Word2Vec による素性追加

2-2 Word2Vec 素性の利用

本研究では Word2Vec を用い,訓練データとテストデータの作成年差を考慮した意味的な素性を追加することにより分類精度の向上を図る．具体的には，Word2Vec を利用して得られた分散表現を意味的な単語のベクトル表現として利用し，事例 X に J 次元から成る意味的な素性ベクトル(Word2Vec 素性)を追加する．図１において学習した Word2Vec モデルを用い Word2Vec 素性を生成し,これを訓練データとテストデータに用いる．

一般に文書分類に利用される素性ベクトルは，各文書データに含まれる単語とその出現回数を Bag-of-Words を用いて表現する．すなわち，各単語とその出現回数で各データをベクトルで表現し，事例間の類似度計算あるいは，機械学習における学習とテストに用いられる．この素性ベクトルを Bag-of-Words 素性とする．例えば，1995 年に報道された野球に関する記事と，2010 年に報道された野球に関する記事文書は，意味的に類似していると考えられる．しかし，選手名や試合会場の名前，試合内容などに関して，単語出現傾向の異なりにより単語の出現回数のみで素性ベクトルを構築した場合，類似した文書と判断されない

場合がある．この問題に対処するため，文書に出現する各単語を意味的に表現する Word2Vec によるベクトル表現を利用し，文書自体の意味を捉えた素性を追加する手法を提案する．先ず，対象とする事例空間に含まれる全ての単語を *W = w1, w2, … wn* とする．*n* は全語彙数である．次に Word2Vec を用い W に関して *J* 次元ベクトルを獲得する．この単語ベクトルを利用し，訓練データ，及びテストデータを表現する．語彙数が I の事例 *X* について，出現する単語を *w=w1, w2,…, wI* とし，各単語の出現回数を *ni* とする．本手法ではこの各単語の出現回数の素性に加え Word2Vec により得られた素性を追加する．追加する J 次元の素性 X を式(1)で示す．

$$x_j = \frac{\sum_{i=1}^{I} w_j^i}{\sum_{i=1}^{I} n_i} \qquad (1)$$

式(1)における分母は，Word2Vec 素性値が事例 X の長さに依存することを防ぐために正規化を行う．式(1)で得られた素性を図 2 で示すように I 次元の Bag-of-words 素性で表現された事例に追加することにより文書分類精度の向上を図る．

2-3 作成年差を考慮した学習

本研究は，訓練データとテストデータの作成年差を考慮した転移学習手法を提案する．一般に，作成年が変化するにつれて単語の出現傾向が異なるため，訓練データとテストデータの作成年が異なる場合，教師付き学習を用いた分類精度は低下する．しかし，訓練データの中には，テストデータと作成年が異なるにもかかわらずテストデータの分類精度に貢献するデータも含まれていると考えられる．したがってそのデータを最大限利用することができれば分類精度を向上させることができると考えられる．本研究は，転移学習を用いて，分類精度の向上を試みる．転移学習とは，解決したい目標問題がデータや知識の不足により解決できない場合，その問題をデータや知識が十分な別の問題のデータを最大限再利用することで解決することを目指す学習手法である．本研究では転移学習を以下のように適用する．

- テストデータと作成年が同じ少量のデータを用いることにより，テストデータと作成年が異なる大量のデータの中から，テストデータの分類精度に貢献するデータを効率良く獲得し高精度な分類を行う．

本研究で用いる転移学習は Dai らが提案した TrAdaBoost を拡張した学習手法である．
TrAdaBoost は，Boosting 手法に基づく分類手法であり，テストデータとは異なる分野が付与された訓練データからテストデータの分類に貢献するデータのみを効率よく利用し分類精度の向上を試みた研究であり，分類に貢献すると判断されたデータには高い重みを付与し，貢献しないと判断されたデータには低い重みを付与して学習を繰り返す方法である．本研究はこの手法を拡張することにより学習を行う．学習手順を以下に示す．

1. 入力と重みベクトルの初期化

分類器を Learner と記す．*n, m* はそれぞれ *Td, Ts* のデータ数である．*Td* の各要素は *xi*(*i=1,…n*)，*Ts* は *xi*(*i=n+1,…,n+m*)である．*Xs* は same 訓練データ，*Xd* は diff 訓練データの各データである．関数 twf(z)は訓練データとテストデータの作成年 *z* ごとに重み *w* = (*w,…,wn+m*)の変化量を決定する関数とする．我々は正規分布を仮定し，twf(z)を定めた．twf(z)を図 3 に示す．

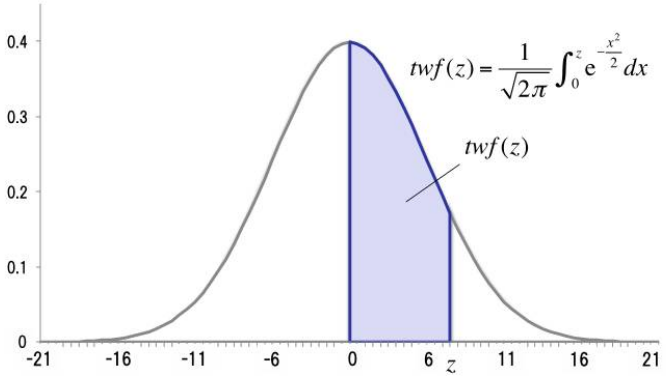


図 3 twf(z)関数

2. 重みベクトル更新の繰り返し：3 から 7 までを *N*回(*t=1, …, N*)繰り返す.
3. 正規化: 式(2)を用い正規化を行い，訓練データ T に重みを積算する.

$$\mathbf{p}^t = \frac{\mathbf{w}^t}{\sum_{i=1}^{n+m} w_i^t} \qquad (2)$$

4. Learner による学習: 訓練データ T とそれに対応する式(2)で求めた重みを Learner に与え学習を行う.
5. 新しい学習器のラベル予測：重みを更新した新しい学習器 *ht* を用い，訓練データ *X* のラベル *Y* を予測する.
6. 誤差の算出：訓練データ *Ts* の学習器 *ht* による誤差 ε_t を式(3)で求める．さらに訓練データ *T* での誤差 δ_t を式(4)で求める.

$$\varepsilon_t = \sum_{i=n+1}^{n+m} \frac{w_i^t I(h_t(x_i) \neq c(x_i))}{\sum_{j=n+1}^{n+m} w_j^t} \qquad (3)$$

$$\delta_t = \sum_{i=1}^{n+m} \frac{w_i^t I(h_t(x_i) \neq c(x_i))}{\sum_{j=1}^{n+m} w_j^t} \qquad (4)$$

ここで *c*(*xi*)は *xi* の正解ラベルを示す．関数 I は 1，または 0 を返す.

7. 重みベクトルの更新： α_t を式(5)で定義し，重みを更新する.

$$\alpha_t = \frac{\log(1-\delta_t/\delta_t)}{2} \qquad (5)$$

➢ diff 訓練データが誤りの場合

$$w_i^{t+1} = w^t \exp(-twf(z)) \tag{6}$$

➢ diff 訓練データが正解の場合

$$w_i^{t+1} = w^t \exp(-\alpha_t h_t(x_i) c(x_i)) \tag{7}$$

8. 出力：βt を式(8)で定義する．

$$\beta_t = \frac{\log(1-\varepsilon_t/\varepsilon_t)}{2} \tag{8}$$

処理を N回繰り返した後，式(9)を用い獲得された学習器による重み付き多数決を行う．

$$F(\mathbf{x}) = \begin{cases} 1, \sum_{t=1}^{N} \beta_t h_t(\mathbf{x}) \geq 0 \\ -1, \ otherwise \end{cases} \tag{9}$$

3. 実験

3-1. データ及び評価方法

本研究における実験では，1991 年から 2012 年までの 22 年分の毎日新聞記事データの中から 8 カテゴリを使用した．カテゴリと総データ数を表 4 に示す．

使用カテゴリ	データ数
国際	178,741
経済	190,672
家庭	91,803
文化	38,092
読書	34,346
芸能	53,378
スポーツ	362,119
社会	562,045
合計	1,511,196

表4：カテゴリと総データ数

	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	合計
T_d	4,062	4,333	2,086	865	780	1,213	8,230	12,773	34,343
T_s	203	216	104	43	39	60	411	638	1,713
S_{para}	1,930	2,059	992	412	371	577	3,910	6,068	16,319
S_{test}	1,930	2,058	991	411	371	576	3,909	6,068	16,315

表 5：実験データ

毎日新聞記事データを diff 訓練データ, same 訓練データ, テストデータに分割する. 分割は毎日新聞記事データを等分に 2 分割し, 一方を diff 訓練データ, 他方をテストデータとし, テストデータの 5%を same 訓練データとした. データとして取り出した. さらにテストデータを 2 分割して一方は各種パラメータ推定用データとし, もう一方は本テストデータとした. 分類実験に用いたデータの各年の平均数を表 5 に示す. 表中の *Spara* はパラメータ推定用テストデータを示し, *Stest* は本テストデータを示す. *Spara* はパラメータの推定にのみ使用し, 本稿に示される S は全て *Stest* である. Word2Vec の素性の次元は 200 次元とした.

提案手法, 及び比較手法として LinerSVM を使用した. 実験では提案手法, 及び Boosting と TrAdaBoost の繰り返し回数, さらに提案手法の重み付け関数 twf(z)の作成年差と標準正規分布の範囲はパラメータ推定用のデータを用い, 最適な数値を選択した. 評価手法には F 値を用いた.

提案手法の有効性を検証するために以下の手法との比較を行った.

(1) SVM: *Td* を用いた SVM
(2) SVMt: *Td* と *Ts* を用いた SVM
(3) Boosting: *Td* と *Ts* を用いた Boosting
(4) TrAdaBoost: *Td* と *Ts* を用いた Dai らの手法

3-2 実験結果

提案手法の有効性を検証するため, diff 訓練データとテストデータの作成年差が± 15, ± 10, ± 5 年における分類精度を求めた. それぞれの結果を表 6, 7, 及び表 8 に示す.

作成年差 +15									
分類手法	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	平均
SVM	0.651	0.661	0.507	0.231	0.621	0.255	0.817	0.758	0.563
SVMt	0.716	0.740	0.658	0.390	**0.845**	0.485	0.912	0.822	0.696
Boosting	0.726	0.745	0.630	0.352	0.827	0.461	0.911	0.834	0.686
TrAdaBoost	0.630	0.661	0.551	0.394	0.759	0.383	0.831	0.810	0.627
提案手法	**0.750**	**0.756**	**0.663**	**0.406**	0.837	**0.543**	**0.923**	**0.853**	**0.716**
作成年差 -15									
SVM	0.619	0.581	0.464	0.269	0.431	0.272	0.835	0.788	0.532
SVMt	0.720	0.728	0.611	0.423	0.689	0.424	0.878	0.837	0.664
Boosting	0.767	0.754	0.605	0.401	0.650	0.452	0.902	0.855	0.673
TrAdaBoost	0.759	0.784	0.587	0.402	0.611	**0.562**	0.881	0.806	0.674
提案手法	**0.793**	**0.806**	**0.662**	**0.483**	**0.690**	0.548	**0.911**	**0.866**	**0.720**

表 6：作成年差±15 における分類精度

作成年差 +10									
分類手法	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	平均
SVM	0.682	0.697	0.497	0.349	0.616	0.383	0.858	0.798	0.610
SVMt	0.733	0.751	0.668	0.510	**0.792**	0.567	0.908	0.840	0.721
Boosting	0.748	0.758	0.642	0.445	0.758	0.547	0.907	0.849	0.707
TrAdaBoost	0.664	0.650	0.420	0.430	0.696	0.260	0.838	0.821	0.597
提案手法	**0.762**	**0.776**	**0.686**	**0.512**	0.769	**0.594**	**0.918**	**0.867**	**0.735**
作成年差 -10									
SVM	0.647	0.644	0.498	0.343	0.490	0.346	0.864	0.802	0.579
SVMt	0.727	0.753	0.642	0.485	0.703	0.489	0.889	0.842	0.691
Boosting	0.772	0.772	0.631	0.459	0.653	0.517	0.905	0.855	0.696
TrAdaBoost	0.738	0.779	0.579	0.261	0.644	0.530	0.880	0.812	0.653
提案手法	**0.788**	**0.810**	**0.678**	**0.532**	**0.713**	**0.598**	**0.908**	**0.867**	**0.737**

表 7：作成年差±10 における分類精度

作成年差 +5									
分類手法	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	平均
SVM	0.725	0.744	0.529	0.398	0.701	0.497	0.880	0.824	0.662
SVMt	0.758	0.790	0.684	0.524	**0.805**	0.615	0.915	0.854	0.743
Boosting	0.774	0.801	0.658	0.488	0.785	0.605	0.914	0.864	0.736
TrAdaBoost	0.686	0.698	0.432	0.409	0.684	0.345	0.875	0.820	0.619
提案手法	**0.784**	**0.807**	**0.689**	**0.525**	0.784	**0.637**	**0.918**	**0.870**	**0.752**
作成年差 -5									
SVM	0.713	0.704	0.533	0.413	0.656	0.464	0.890	0.832	0.651
SVMt	0.756	0.778	0.673	0.529	**0.777**	0.576	0.915	0.854	0.743
Boosting	0.789	0.790	0.659	0.500	0.731	0.592	0.919	0.873	0.732
TrAdaBoost	0.724	0.753	0.418	0.243	0.647	0.518	0.877	0.824	0.626
提案手法	**0.798**	**0.813**	**0.690**	**0.534**	0.763	**0.645**	**0.919**	**0.878**	**0.755**

表 8：作成年差±5 における分類精度

表の各値は，各手法の分野ごとの F 値を示し，平均はマクロ F 値を示す．表 6，7，及び 8 より本手法は，各分野において他手法よりも高い精度が得られていることがわかる．一方，提案手法は「読書」カテゴリで SVMt よりも精度が低下している．「読書」カテゴリは，最もデータ数の少ないカテゴリであり，Boosting でも「読書」カテゴリの精度が低下していることから，繰り返し処理により少量の正例カテゴリである「読書」カテゴリの文書の重みが荷重になり，結果的に正解率が低下すると同時に F 値が低下したことが考えられる．また，比較手法の SVMt と Boosting を比較すると，作成年差が－15 のときは Boosting の精度が高いものの＋15 年のときは SVMt よりも精度が低下している．これは Boosting は誤分類した diff 訓練データも高い重みを付与するが，テストデータとの作成年差を考慮することができないため，分類に不要なデータの重みを増加させてしまった結果，分類精度が低下したことが考えられる．一方，TrAdaBoost の結果は，作成年差＋15 年で SVMt よりも低下している．これは TrAdaBoost における重みにおいて，diff 訓練データの重みを下げすぎてしまったことが原因として考えられる．
表より，本手法は，他手法と比較した場合，特に年度差が大きい場合に，精度の大きな向上がみられると言える．

3-3 作成年差による精度

次に，毎日新聞記事データの 1991 年から 2012 年までを使用し，Td と Ts，及び S の作成年差による精度を求めた．実験結果を図 4 に示す．

図 4 において縦軸は F 値を示し，横軸は訓練データとテストデータの作成年差を示す．図 4 より，本手法は，SVMt，Boosting，及び TrAdaBoost と比較しいずれの作成年差においても優れていることがわかる．特に SVMt と比較すると，SVMt が訓練データとテストデータとの作成年差が大きくなるほど本手法は，精度が向上している．SVMt は、テストデータが訓練データよりも後で作成された場合に精度がよいことがわかる．このことは，テストデータが訓練データよりも後で作成された場合に、より分類が難しいことを示している．一方，本手法は，テストデータが訓練データよりも早期に作成された場合，すなわち，横軸がマイナスのときに高精度な分類が可能となっている．このことから，本手法は作成年差の正負に関わらず，比較手法の精度を向上させることができていると言える．

一方，作成年差が±1 のときに提案手法は，SVMt よりもわずかに精度が低下している．作成年差が±１のときは，テストデータの分類に有効な diff 訓練データが多数存在すると考えられる．本手法で用いている正規分布は作成年差に基づいており，訓練データの量は反映していない．そのため，誤分類した場合，テストデータに有効な diff 訓練データの重みを低下させてしまい，結果的に分類精度を低下させていることが考えられる．今後は，訓練データの量も考慮した重み付け関数を検討する必要がある．

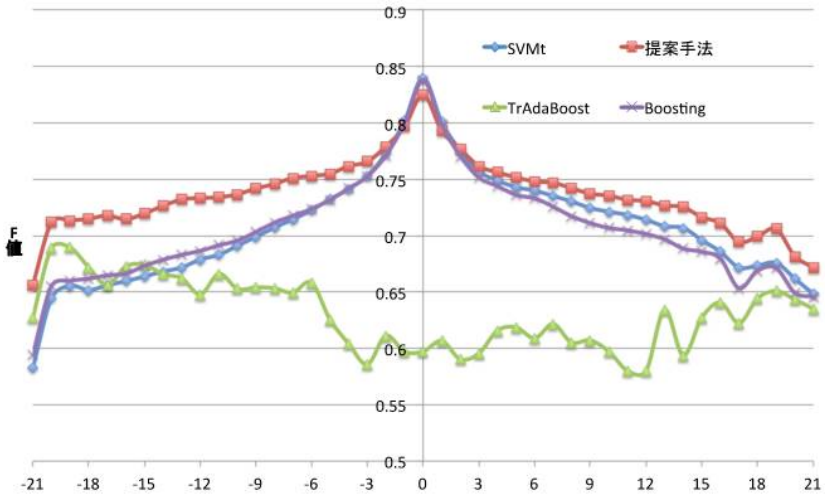


図 4: 作成年差による分類精度

3-4. Word2Vec 素性の有効性

本節では, Word2Vec 素性の追加が, 作成年が異なる文書の分類において精度に貢献するかを調査した. 作成年差が＋15 年と－15 年における結果を表 9 に示す.

表 9 は, Bag-of-words 素性のみで分類を行った場合と Bag-of-words 素性に Word2Vec 素性を導入した SVMt, 及び提案手法との比較を示す. 実験結果から提案手法においてほぼ全てのカテゴリにおいて Word2Vec 素性を追加した場合に精度が向上していることがわかる. SVMt においても「国際」カテゴリ以外のカテゴリで精度が向上している. このことから Word2Vec 素性は訓練データとテストデータの作成年が異なる場合に有効であると言える. SVMt と本手法とを比較すると, 本手法の精度が優れており, 特に－15 年における分類に貢献していることがわかる.

3-5. Word2Vec における次元数の効果

Word2Vec の次元数が与える精度の影響について調査するため, 作成年差を±15 年とし, SVMt と提案手法で比較を行った. 次元数は 100 次元から 900 次元まで 100 ステップずつ変化させ精度を求めた. 実験結果を図 5 に示す.

SVMt									
分類手法	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	平均
BoW	**0.722**	0.729	0.641	0.352	0.806	0.452	0.904	0.818	0.678
BoW+w2v	0.716	**0.741**	**0.658**	**0.390**	**0.844**	**0.484**	**0.912**	**0.822**	**0.696**
提案手法									
BoW	0.725	0.731	0.637	0.388	0.830	0.509	0.905	0.842	0.696
BoW+w2v	**0.749**	**0.756**	**0.663**	**0.407**	**0.838**	**0.542**	**0.923**	**0.853**	**0.716**

SVMt									
分類手法	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	平均
BoW	0.693	0.705	0.591	0.403	0.640	0.410	0.872	0.829	0.643
BoW+w2v	**0.720**	**0.728**	**0.611**	**0.426**	**0.688**	**0.423**	**0.879**	**0.837**	**0.664**
提案手法									
BoW	0.773	0.781	0.621	0.460	0.653	0.515	0.891	0.857	0.694
BoW+w2v	**0.793**	**0.805**	**0.662**	**0.483**	**0.690**	**0.545**	**0.911**	**0.866**	**0.719**

表 9: Word2Vec 素性による分類精度

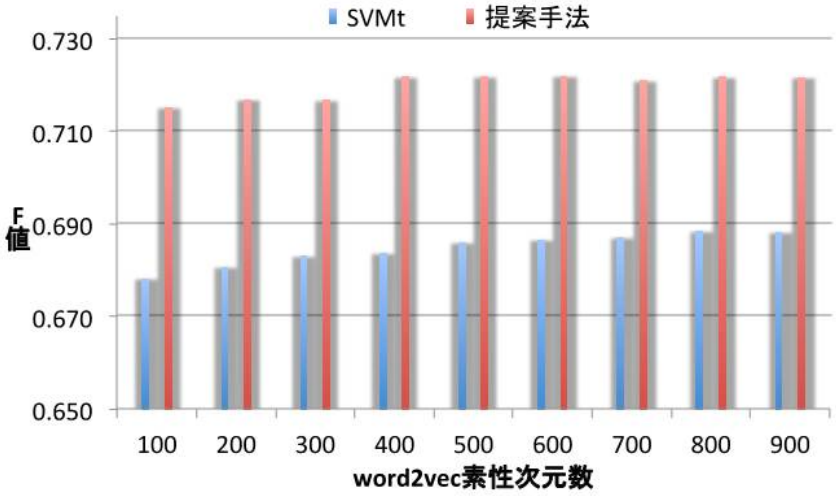


図 5: Word2Vec 素性次元数と分類精度

図 5 より提案手法は 100 次元から 300 次元まではほとんど精度が変化していない．一方 400 次元において精度の向上がみられ，それ以降は次元数が増加しても精度はほとんど変化しなかった．同様に，SVMt の精度も次元数が 100 威厳のときに最も精度が低かった．500 次元までは次元数の増加と共に精度も上昇し 800 次元で最も高い精度が得られた．このことから，Word2Vec の素性次元数の推定は分類精度に影響を与えることがわかる．

3-6 単語スムージングの効果

Word2Vec を利用した単語のスムージングの効果を検証するため，スムージングの有無による実験を行った．作成年差±15 年の場合における実験結果を表に示す．実験では Word2Vec 素性による効果を排除するために Bag-of-words 素性のみを用いた実験を行った．結果を表 10，及び 11 に示す．

SVMt									
分類手法	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	平均
スムージング無	**0.7219**	0.7285	**0.6424**	**0.3534**	0.8058	0.4495	0.9041	**0.8183**	0.6780
スムージング有	0.7217	**0.7286**	0.6413	0.3521	**0.8062**	**0.4519**	**0.9043**	0.8182	0.6780
提案手法									
スムージング無	0.7247	0.7307	0.6370	0.3878	0.8302	0.5091	**0.9053**	0.8422	0.6959
スムージング有	**0.7255**	**0.7341**	**0.6376**	**0.4028**	**0.8330**	**0.5165**	0.9052	**0.8437**	**0.6998**

表 10: 作成年差が＋15 年における分類精度

SVMt									
分類手法	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	平均
スムージング無	0.6929	0.7046	0.5905	**0.4032**	0.6400	**0.4099**	0.8724	**0.8293**	0.6429
スムージング有	**0.6938**	**0.7056**	**0.5909**	0.4017	**0.6415**	0.4098	**0.8728**	0.8291	**0.6431**
提案手法									
スムージング無	**0.7732**	**0.7813**	0.6209	**0.4595**	**0.6526**	**0.5145**	**0.8912**	**0.8570**	**0.6938**
スムージング有	0.7661	0.7803	**0.6229**	0.4544	0.6491	0.5049	0.8896	0.8534	0.6901

表 11: 作成年差が－15 年における分類精度

表の上段は作成年差が＋15 年における分類精度を示し下段は－15 年における精度を示す．表 10，及び 11 より，いずれの分野においてもスムージング有において無の場合と比較し有意な差が得られなかった．さらに提案手法では＋15 年の場合に，わずかに精度の向上が見られたが，－15 年の場合には，わずかに精度が低下していることがわかる．そこで，スムージングにより未知語が訓練データに存在する類似語に置き換わったかを調査した．

類似した単語で置き換えられた例		類似しない単語で置き換えられた例	
スムージング前	スムージング後	スムージング前	スムージング後
多恵子	多恵子	狭間が丘	岡田
中小公庫	日銀	サガ	アテる
三菱鉱業セメント	三菱金属鉱業	長刀	佳和

表 12: スムージング結果例

表 12 は，スムージングの結果を示す．表より，”中小公庫”と“日銀”，“三菱鉱業セメント”と”三菱金属鉱業”は同じカテゴリに出現することの多い類似単語で置き換わっていることがわかる．一方，“多恵子”と“多恵子”は異なる人名を表しており出現するカテゴリも異なる可能性がある単語であるが置き換わっている．50 単語を無作為に抽出して調査した結果，26 単語は意味的に類似していない単語であった．単語のスムージングについては，さらなる検討が必要である．

3-7 重み付け関数の有効性

重み付け関数の有効性を検証するために，正規分布の入力の範囲を[-X，+X]とし，作成年差－21 年から＋21 年を割り当てた場合の精度を表 13 に示す．単語スムージングと同様，Word2Vec とスムージングによる効果を排除し重み付け関数による効果を検証するため Bag-of-words 素性のみを用いて実験を行った．表より，本手法は，重み付け関数の入力範囲を変更した場合においても比較手法である SVMt よりも高い精度が得られている．SVMt は作成年差が＋15 年において[-0.5,0.5]の時に最高精度であり－15 年においては[-0.7,0.7]のときに最高精度が得られた．SVMt は，diff 訓練データとテストデータとの単語の出現傾向の変化は－15 年の方が大きいと考えられる．このことから，＋15 年では重みの極端な低下が少ない範囲が有効であり－15 年では大きな範囲が有効であると言える．

作成年差 +15									
分類手法	国際	経済	家庭	文化	読書	芸能	スポーツ	社会	平均
SVMt	0.722	0.729	0.642	0.353	0.806	0.450	0.904	0.818	0.678
[-0.3,0.3]	0.724	0.750	0.647	0.368	0.835	0.493	0.910	0.845	0.697
[-0.5,0.5]	0.730	0.741	0.644	0.395	0.841	0.511	0.910	0.846	**0.702**
[-0.7,0.7]	0.725	0.731	0.637	0.388	0.830	0.509	0.905	0.842	0.696
[-1.0,1.0]	0.723	0.729	0.636	0.373	0.826	0.512	0.904	0.839	0.693
作成年差 -15									
SVMt	0.693	0.705	0.590	0.403	0.640	0.410	0.872	0.829	0.643
[-0.3,0.3]	0.764	0.769	0.617	0.442	0.648	0.494	0.891	0.854	0.685
[-0.5,0.5]	0.759	0.774	0.617	0.448	0.657	0.499	0.887	0.853	0.687
[-0.7,0.7]	0.773	0.781	0.621	0.459	0.653	0.514	0.891	0.857	**0.694**
[-1.0,1.0]	0.768	0.785	0.622	0.443	0.650	0.515	0.888	0.848	0.690

表 13: 重み付け関数による分類精度

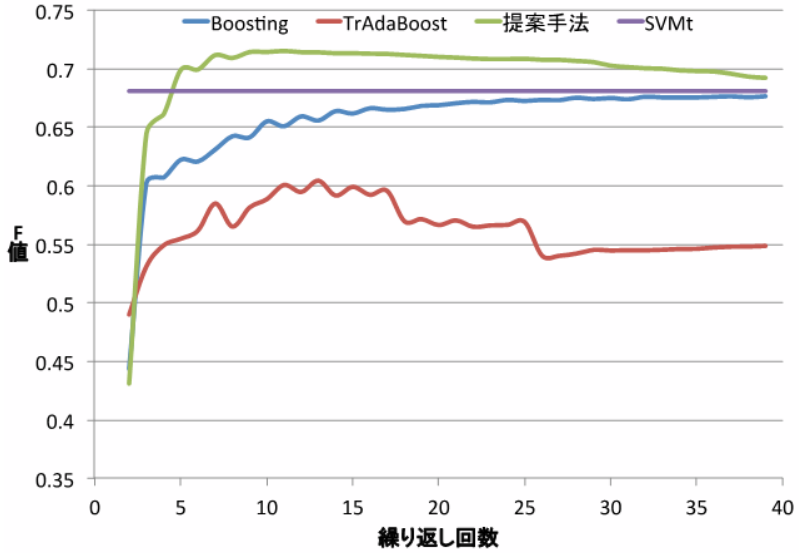


図 6: 繰り返し数による分類精度

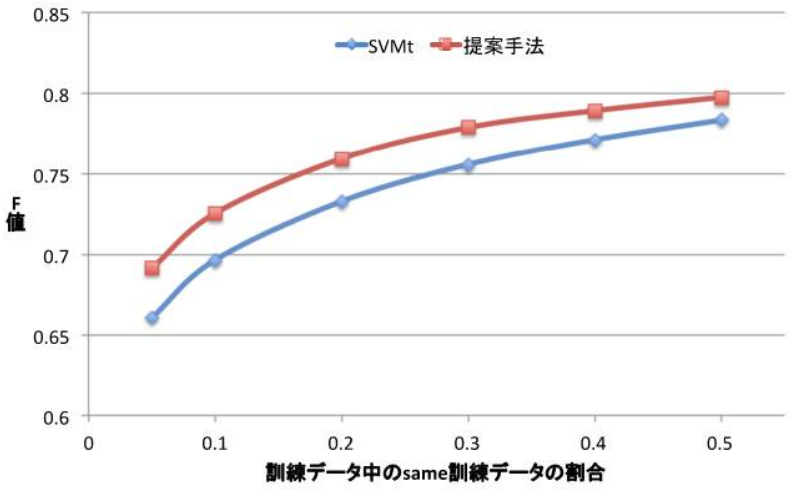
3-8 繰り返し回数による分類精度

Boosting を用いた手法の精度は，繰り返し数に依存する．したがって，適切な繰り返し数の推定が精度に貢献する．そこで，各手法について，繰り返し数と精度を求めた．実験結果を図 6 に示す．

図 6 は，繰り返し回数を 2 回から 40 回とした場合における各手法の精度を示す．図の横軸は繰り返し回数を示し，縦軸は F 値を示す．図 6 より，本手法は繰り返し回数が 10 回のときに最高精度を示し，早い段階で SVMt よりも高い精度で安定していることがわかる．一方，繰り返し回数が 10 回以上になると，精度が徐々に低下している．これは重み付けを過剰に行っているためであり，diff 訓練データと same 訓練データとの重みが偏っているためであると考えられる．

3-9 same 訓練データの割合と分類精度

訓練データに占める same 訓練データの割合が多いほど，テストデータを高精度で分類することが可能となる．しかし same 訓練データ数は限られているため，diff 訓練データから分類に有効なデータを如何に抽出するかが重要となる．本節では、same 訓練データの割合が分類精度に与える影響を調べるため，割合を変化させた場合の精度を求めた．実験結果を図 7 に示す．



,

図 7: same 訓練データの割合と分類精度

図 7 は，作成年差が±15 年の結果を示し，全訓練データ中に含まれる same 訓練データの割合と精度との関係を示す．横軸は訓練データ全体に対する same 訓練データの割合を示し，縦軸はマクロ F 値を示す．図 7 より same 訓練データの割合にかかわらず提案手法が SVMt よりも高い精度が得られていることがわかる．same 訓練データの割合が 50%のときは，SVMt と提案手法との精度差は約 2%であるのに対し，same 訓練データの割合が 5%になると約 4%になっていることから，本手法は SVMt と比較し same 訓練データが少ない場合にも高い精度で分類できることが確認できた．

3. Web ディレクトリの自動改変

3-1. ディレクトリの統合による自動改変

本研究では，2 つのディレクトリを対象とし，双方におけるディレクトリの整合性を保持しつつ，これらを統合することにより，新たな分野を推定する手法を提案する．提案手法は，(1)類似カテゴリ対の抽出，(2)カテゴリの合併，(3)カテゴリの割り当ての 3 つのオペレーションから成る．図 8 に提案手法の流れを示す．

(1) 類似カテゴリ対の抽出

図 8 において H_A, H_B から類似しているカテゴリ組を抽出する．H_A に属する文書を H_B に分類し，各カテゴリに分類された文書の数からカテゴリ間の類似性を推定することによりカテゴリ組を抽出する．分類には機械学習 SVM を用いた．式(11)は A と B における類似度尺度を示す．

$$X^2(A,B) = \begin{cases} 0 & (|ad-bc| \le N/2) \\ \frac{N\times(|ad-bc|-N/2)^2}{e\times f\times g\times h} & (otherwise) \end{cases} \qquad (10)$$

式において，a, b, c, d はそれぞれカテゴリ A に分類された文書が B にも分類される文書数，A に分類された文書が B に分類されない文書数，A 以外に分類された文書が B に分類される文書数，A 以外に分類された文書が B 以外に分類される文書数を示す．N は総文書数を示す．また $e=a+b$, $f=c+d$, $g=a+c$, $h=b+d$ を示す．式の値が大きいほど，カテゴリ A と B は類似していることを示す．式(10)は，カテゴリ A，及び B に属する文書数に依存するため式(11)を用いて正規化を行う．

$$X^2_{new} = \frac{X^2_{old} - X^2_{min}}{X^2_{max} - X^2_{min}} \qquad (11)$$

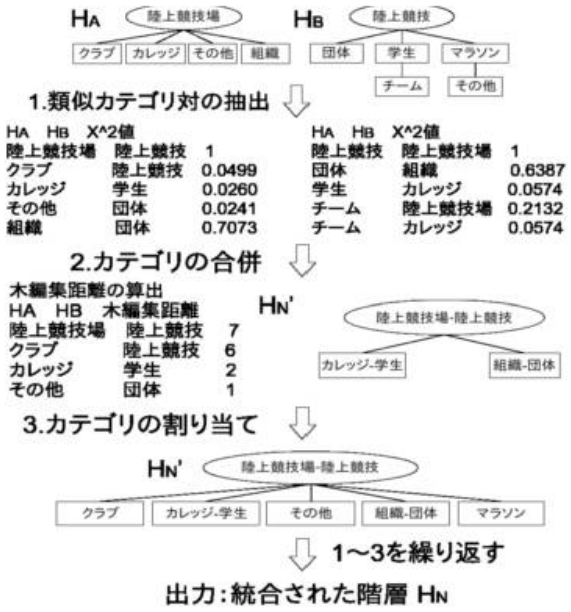


図 8: 処理の流れ

(2) カテゴリの合併

(1) で得られた類似カテゴリの組について，以下の処理によりカテゴリ同士の合併を行う．

(a) H_A, H_B に同名のカテゴリが存在していた場合，同名のカテゴリ同士を合併する．

(b) 同名でないカテゴリは，(1)の類似カテゴリにより抽出されたカテゴリ組を利用する．具体的には H_A に属する文書を H_B に分類した結果得られたカテゴリ組と H_B に属する文書を H_A に分類した結果得られたカテゴリ組の両方に属するカテゴリ組を合併する．

(c) 一つのカテゴリに対して複数のカテゴリ対が抽出されていた場合，木編集距離が小さいカテゴリ組を合併する．ここで，木編集距離とは，異なる 2 つの木を同形にするための操作コストの総和であり，木編集距離が小さいほど 2 つの木は類似していると判断される．同形にするための操作として(i) 頂点の削除，(ii) 頂点の挿入，(iii) 頂点の値の変更の 3 つがあり，それぞれの操作のコストは 1 である．また 3 つの操作において，操作の前後で任意の頂点間の順序が変更されることはない．

(3) カテゴリの割り当て

処理(2)カテゴリの合併により合併されないと判断されたカテゴリを階層に割り当てる．各カテゴリの親子関係は保持される必要があるため，統合前の階層構造における親子関係を利用する．割り当ての対象となるカテゴリの親カテゴリが合併されていた場合，合併されたカテゴリの子として割り当てる．

3-2. 実験

3-2-1. データ及び評価方法

実験では DMOZ の「スポーツ」と Yahoo の「スポーツ」を利用した．それぞれの階層において，登録されている Web サイト内に存在する他ページへのリンクを最大 2 回辿ることで Web ページを取得し，各カテゴリの文書とした．それぞれの階層において，1,000 以上の文書が存在するカテゴリを使用した．また，第 1 階層のカテゴリを文書が多い順にソートし，合併されるであろうと考えられるカテゴリを 7 カテゴリずつ，そうでないカテゴリを 2 カテゴリずつ使用した．また，各カテゴリに無作為に 12,000 文書を取得し，10,000 文書を訓練データ，2,000 文書をテストデータとした．DMOZ の「格闘技」に属する文書数は少量であったため， 879 文書をテストデータとした．実験で使用したカテゴリを表 14 に示す．丸括弧内にカテゴリ数を示す．Yahoo のカテゴリは，翻訳後のカテゴリ名を示す。また，実験に使用したカテゴリの一部を図 9 に示す． /の左側のカテゴリを親カテゴリ，右側のカテゴリを子カテゴリとする．例えば，「スポーツ/陸上競技」は「スポーツ」の子カテゴリとして「陸上競技」が存在するということを示す．

./スポーツ/陸上競技
./スポーツ/ウォータースポーツ
./スポーツ/ゴルフ
./スポーツ/バレーボール
./スポーツ/野球
./スポーツ/格闘技
./スポーツ/バスケットボール
./スポーツ/武道・武術
./スポーツ/サッカー
./スポーツ/陸上競技/団体
./スポーツ/陸上競技/学生
./スポーツ/陸上競技/マラソン
./スポーツ/ウォータースポーツ/セーリング
./スポーツ/ウォータースポーツ/サーフィン
./スポーツ/ゴルフ/コース
./スポーツ/ゴルフ/用具
./スポーツ/ゴルフ/ガイドとディレクトリ
./スポーツ/バレーボール/選手
./スポーツ/バレーボール/ビーチバレー
./スポーツ/バレーボール/学生
./スポーツ/野球/球場
./スポーツ/野球/選手
./スポーツ/野球/ベースボール・チャレンジ・リーグ
./スポーツ/野球/日本野球機構
./スポーツ/野球/ニュースとメディア
./スポーツ/野球/四国アイランドリーグplus
./スポーツ/野球/審判
./スポーツ/野球/アマチュア
./スポーツ/野球/チーム
./スポーツ/格闘技/その他
./スポーツ/格闘技/レスリング
./スポーツ/格闘技/総合格闘技
./スポーツ/格闘技/ボクシング

図 9: カテゴリ階層例

DMOZ(148)	Yahoo(177)
陸上競技 (6)	陸上競技場 (5)
サッカー (42)	サッカー (42)
野球 (35)	野球 (38)
ゴルフ (11)	ゴルフ (14)
バレーボール (6)	バレーボール (7)
ウォータースポーツ (6)	サーフィン (7)
格闘技 (13)	武道 (31)
武道・武術 (20)	ボクシング (7)
バスケットボール (9)	オートレーシング (26)

表 14: 各ディレクトリにおけるカテゴリ数

本研究における評価尺度は，正解率を用いた．実験では統合を行うことにより，誤分類した文書の数が少なくなるかを調査し，統合の有効性を確認した．実験結果を表 15，及び 16 に示す．

	第 1 階層	第 1-第 2	第 1-第 3	第 1-第 4
DMOZ-Yahoo	97.78	83.98	75.48	66.99
Yahoo-Dmoz	95.29	82.33	71.56	69.99
平均	96.53	83.16	73.52	68.49

表 15: 統合前 :「ゴルフ」での正解率 (%)

	第 1 階層	第 1-第 2	第 1-第 3	第 1-第 4
DMOZ-統合後	99.78	99.66	90.17	83.14
Yahoo-統合後	99.68	99.5	92.44	86.44
平均	99.73	99.58	91.31	84.79

表 16: 統合後 :「ゴルフ」での正解率 (%)

表 15，及び 16 は、統合前に DMOZ の「ゴルフ」を Yahoo に分類した結果と Yahoo の「ゴルフ」を DMOS に分類した結果を示す。表は，統合後の各々を示す．表 15，及び 16 より，統合前と統合後において，統合後の階層における正解率が高いことがわかる．第 3 行 4 列の値を比較すると，統合後の階層では正解率が 16.3%高くなっている．DMOZ の文書を Yahoo に分類した場合と統合後の階層に分類した場合，Yahoo の文書を DMOZ に分類した場合と統合後の階層に分類した場合のどちらにおいても正解率が高くなることが確認できた．また，統合前の階層では階層が深くなるにつれて正解率が大きく下がっている．しかし統合後の階層では統合前の階層ほど正解率の下がり幅が小さいことから，階層の統合の有効性を確認することができた．

次に統合後のカテゴリ数の調査を行った．結果を表 17 に示す．

DMOZ(148)	Yahoo(177)	統合した階層 (292)
陸上競技 (6)	陸上競技場 (5)	陸上競技-陸上競技場 (8)
サッカー (42)	サッカー (42)	サッカー (77)
野球 (35)	野球 (38)	野球 (62)
ゴルフ (11)	ゴルフ (14)	ゴルフ (23)
バレーボール (6)	バレーボール (7)	バレーボール (11)
ウォータースポーツ (6)	サーフィン (7)	ウォータースポーツ (10)
格闘技 (13)	武道 (31)	格闘技-武道 (46)
武道・武術 (20)	ボクシング (7)	武道・武術 (20)
バスケットボール (9)	オートレーシング (26)	バスケットボール (9)
		オートレーシング (26)

表 17: 統合後のカテゴリ数

表 17 は，統合前のカテゴリ数と統合後のカテゴリ数を示す．丸括弧内にカテゴリ数を示す． 表 17 より，統合後の階層では総カテゴリ数が 292 と大きく増加していることがわかる．そのため，統合後の階層に対して，さらにカテゴリ同士を合併する手法について今後検討する必要がある．

4. まとめ

本研究は，既存の分類体系に合致しない文書について，新たな分野を自動的に推定する手法を提案した．具体的には，Web ページ文書を分類するために，(1) 時間差適応を行う分類モデルを提案した．さらに，(2) 2 つのディレクトリを対象とし，双方におけるディレクトリの整合性を保持しつつ，これらを統合することにより，新たな分野を推定する手法を提案した． 今後の課題として，(1) 少量の訓練データに対する精度向上，(2) 大規模階層構造を用いた定量的な評価について検討する必要がある．

【参考文献】

[1] T. Mikolov, K. Chen, G. Corrado, J. Dean and J. Dean, Efficient Estimation of Word Representations in Vector Space, http://arxiv.org/abs/1301.3781, 2013.

[2] D. Xing, W. Dai, G-R. Xue and Y. Yu, Bridged Refinement for Transfer Learning, In Proc. of the 11th European Conference on Principles of Data Mining and Knowledge discovery, pp. 324-335, 2007.

[3] L. Breiman, Bagging Predictors, In Machine Learning, Vol. 24, Issue 2, pp. 124-140. 1996.

[4] W. Dai, G. Xue, Q. Yang and Y. Yu, Co-clustering based Classification for Out-of-Domain Documents, In Proc. of the 13th International Conference on Knowledge discovery and Data Mining, pp. 210-219, 2007.

[5] W. Dai, Q. Yang, G-R. Xue and Y. Yu., Boosting for Transfer Learning, In Proc. of the 14th International Conference on Machine Learning, pp. 193-200, 2007.

[6] T. Salles, L. Rocha, G. L. Pappa, F. Mourao, W. J. Meria and M. Goncalves, Temporally-Aware Algorithms for Document Classification, In Proc. of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 307-314, 2010.

[7] X. Chao, D. Wang, Z. Zhang and C. Liu, Document Classification with Distributions of Word Vectors, In Asia-Pacific Signal and Information Processing Association 2014, Annual Summit and Conference, pp. 1-4, 2014.

[8] Q. Le and T. Mikolov, Distributed Representations of Sentences and Documents, in arXiv Preprint arXiv:1405.4053, 2014.

[9] G. Long, L. Chen, X. Zhu and C. Zhang, TCSST: Transfer Classification of Short & Sparse Text Using External Data, In Proc. of the 21st ACM International Conference on Information and Knowledge Management.
[10] N. Chos, I. Song and H. Han, A Survey on Ontology Mapping, Newsletter ACM SIGMOD Record, Vol. 3, No. 35, pp. 34-41, 1999.
[11] L. He and X. Sun, Automatic Maintenance of the Category Hierarchy, In Proc. of the 9th International Conference on Semantics, Knowledge and Grids, pp. 218-221, 2013.
[12] M. Kusner, Y. Sun, N. I. Kolkin and K. Q. Weinberger, From Word Embeddings to Document Distances, In Proc. of the 32nd International Conference on Machine Learning, pp. 957-966.
[13] M. Pawlik and N. Augsten, DRTED: A robust Algorithm for the Tree Edit Distance, In Proc. of the 38th International Conference on Very Large Data Bases, pp. 334-345, 2012.
[14] Q. Yuan, G. Cong, A. Sun, C-Y. Lin and N. M. Thalmann, Category Hierarchy Maintenance: A Data-Driven Approach, In Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 791-800, 2012.
[15] J. Zhang and J. Zhang and S. Chen, Constructing Dynamic Category Hierarchies for Novel Visual Category discovery, In Proc. of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 479-491, 2012.
[16] H. Zhuge and L. He, Automatic Maintenance of Category Hierarchy, Future Generation Computer Systems, Vol. 67, No. 1, pp. 1-12, 2017.

〈発 表 資 料〉

題　　名	掲載誌・学会名等	発表年月
Identification of Domain-Specific Senses based on Word Embedding Learning, F. Fukumoto and Y. Suzuki	Proc. of the 8th Language and Technology Conference, pp. 179-183	2017/11/19
Title Categorization based on Category Granularity, F. Fukumoto and Y. Suzuki	Proc. of the 8th Language and Technology Conference, pp. 83-87	2017/11/19
Is (President,大統領) a Correct Sense Pair? Linking and Creating bilingual Sense Correspondences, F. Fukumoto and Y. Suzuki	Prof. of the 9th International Conference on Knowledge Engineering and Ontology Development, pp. 39-48	2017/11/1
Smoothing Temporal Difference for Text Categorization, F. Fukumoto and Y. Suzuki	Lecture Notes in Computer Science, pp. 203-214	2016/5