Temporal-based Feature Selection and Transfer Learning for Text Categorization

Fumiyo Fukumoto and Yoshimi Suzuki
Graduate Faculty of Interdisciplinary Research, University of Yamanashi, Kofu, Japan

Keywords: Feature Selection, Latent Dirichlet Allocation, Temporal-based Features, Text Categorization, Timeline Adaptation, Transfer Learning.

Abstract: This paper addresses text categorization problem that training data may derive from a different time period from the test data. We present a method for text categorization that minimizes the impact of temporal effects. Like much previous work on text categorization, we used feature selection. We selected two types of informative terms according to corpus statistics. One is temporal independent terms that are salient across full temporal range of training documents. Another is temporal dependent terms which are important for a specific time period. For the training documents represented by independent/dependent terms, we applied boosting based transfer learning to learn accurate model for timeline adaptation. The results using Japanese data showed that the method was comparable to the current state-of-the-art biased-SVM method, as the macro-averaged F-score obtained by our method was 0.688 and that of biased-SVM was 0.671. Moreover, we found that the method is effective, especially when the creation time period of the test data differs greatly from that of the training data.

1 INTRODUCTION

Text categorization supports to improve many tasks such as automatic topic tagging, building topic directory, spam filtering, creating digital libraries, sentiment analysis in user reviews, information retrieval, and even helping users to interact with search engines (Mourao et al., 2008). A growing number of machine learning (ML) techniques have been applied to the text categorization task (Xue et al., 2008; Gopal and Yang, 2010). For reasons of both efficiency and accuracy, feature selection is often used since the early 1990s when applying machine learning methods to text categorization (Lewis and Ringuette, 1994; Yang and Pedersen, 1997; Dumais and Chen, 2000). Each document is represented using a vector of selected features/terms (Yang and Pedersen, 1997; Hassan et al., 2007). Then, the documents with category label are used to train classifiers. Once category models are trained, each test document is classified by using these models. A basic assumption in the categorization task is that the distributions of terms between training and test documents are identical. When the assumption is not hold, the classification accuracy was worse. However, it is often the case that the term distribution in the training data is different from that of the test data when the training data may drive from a different time period from the test data. For instance, the term "Alcindo" frequently appeared in the documents tagged "Sports" category in 1994. This is reasonable because Alcindo is a Brazilian soccer player and he was one of the most loved players in 1994. However, the term did not occur in more frequently in the Sports category since he retired in 1997. The observation show that the informative term appeared in the training data, is not informative in the test data when training data may derive from a different time period from the test data, *e.g.*, in the above example, the term "Alcindo" is informative in the training data with Sports category collected from 1994, but not informative in the test data from other years, *e.g.*, 2005 which should be classified into the Sports category. Moreover, manual annotation of tagged new data is very expensive and time-consuming. The methodology for accurate classification of the new test data by making the maximum use of tagged old data is needed in both feature selection and learning techniques.

In this paper, we present a method for text categorization that minimizes the impact of temporal effects. We selected two types of salient terms by using a simple feature selection technique, χ^2 statistics. One is temporal independent terms that are salient

Fukumoto, F. and Suzuki, Y..
Temporal-based Feature Selection and Transfer Learning for Text Categorization.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 17-26
ISBN: 978-989-758-158-8

across full temporal range of training documents such as “baseball” and “tennis” in the Sports category. Another is temporal dependent terms that are salient for a specific time period such as “Alcindo” in the Sports category in 1994 mentioned in the above example. Hereafter, we call it temporal-based feature selection (TbFS). As the result of TbFS, each document is represented by using a vector of the selected independent/dependent terms, and classifiers are trained. We applied boosting based transfer learning, called TrAdaboost (Dai et al., 2007) in order to minimize the impact of temporal effects. Hereafter, we call it temporal-based transfer learning, TbTL. The idea is to use TrAdaboost to decrease the weights of training instances that are very different from the test data.

The rest of the paper is organized as follows. The next section describes an overview of existing related work. Section 3 presents our approach, especially describes how to adjust temporal difference between training and test documents. Finally, we report some experiments with a discussion of evaluation.

2 RELATED WORK

The analysis of temporal aspects is a practical problem as well as the process of large-scale heterogeneous data since the World-Wide Web (WWW) is widely used by various sorts of people. It is widely studied in many text processing tasks. One attempt is concept or topic drift dealing with temporal effects (Kleinberg, 2002; Lazarescu et al., 2004; Folino et al., 2007). The earliest known approach is the work of (Klinkenberg and Joachims, 2000). They presented a method to handle concept changes with SVMs. They used $\xi\alpha$-estimates to select the window size so that the estimated generalization error on new examples is minimized. The results which were tested on the TREC show that the algorithm achieves a low error rate and selects appropriate window sizes. Wang *et al.* developed the continuous time dynamic topic model (cDTM) (Wang et al., 2008). The cDTM is an extension of the discrete dynamic topic model (dDTM). The dDTM is a powerful model. However, the choice of discretization affects the memory requirements and computational complexity of posterior inference. cDTM replaces the discrete state space model with its continuous generalization, Brownian motion. He *et al.* proposed a method to find bursts, periods of elevated occurrence of events as a dynamic phenomenon instead of focusing on arrival rates (He and Parker, 2010). They used Moving Average Convergence/Divergence (MACD) histogram which was used in technical stock market analysis (Murphy, 1999) to detect bursts. They tested the method using MeSH terms and reported that the model works well for tracking topic bursts. He *et al.* bursts model can be regarded as salient features/terms identification for a specific time period, although their method can not extract such terms automatically, *i.e.* it is necessary to give these terms in advance as the input of their model.

Another attempt is domain adaptation. The goal of this attempt is to develop learning algorithms that can be easily ported from one domain to another, *e.g.*, from newswire to biomedical documents (III, 2007). Domain adaptation is particularly interesting in Natural Language Processing (NLP) because it is often the case that we have a collection of labeled data in one domain but truly desire a model that can work well for another domain. Lots of studies addressed domain adaptation in NLP tasks such as part-of-speech tagging (Siao and Guo, 2013), named-entity (III, 2007), and sentiment classification (Glorot et al., 2011) are presented. One approach to domain adaptation is to use transfer learning. The transfer learning is a learning technique that retains and applies the knowledge learned in one or more tasks to efficiently develop an effective hypothesis for a new task. The earliest discussion is done by ML community in a NIPS-95 workshop[1], and more recently, transfer learning techniques have been successfully applied in many applications. Blitzer *et al.* proposed a method for sentiment classification using structual correspondence learning that makes use of the unlabeled data from the target domain to extract some relevant features that may reduce the difference between the domains (Blitzer et al., 2006). Several authors have attempted to learn classifiers across domains using transfer learning in the text classification task (Raina et al., 2006; Dai et al., 2007; Sparinnapakorn and Kubat, 2007). Raina *et al.* proposed a transfer learning algorithm that constructs an informative Baysian prior for a given text classification task (Raina et al., 2006). The prior encodes useful domain knowledge by capturing underlying dependencies between the parameters. They reported that a 20 to 40% test error reduction over a commonly used prior in the binary text classification task. All of these approaches mentioned above aimed at utilizing a small amount of newly labeled data to leverage the old data to construct a high-quality classification model for the new data. However, the temporal effects are not explicitly incorporated into their models.

To our knowledge, there have been only a few previous works on temporal-based text categorization

[1] http://socrates.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html.

(Kerner et al., 2008; Song et al., 2014). Mourao *et al.* investigated the impact of temporal evolution of document collections based on three factors: (i) the class distribution, (ii) the term distribution, and (iii) the class similarity. They reported that these factors have great influence in the performance of the classifiers throughout the ACM-DL and Medline document collections that span across more than 20 years (Mourao et al., 2008). Salles *et al.* presented an approach to classify documents in scenarios where the method uses information about both the past and the future, and this information may change over time (Salles et al., 2010). They address the drawbacks of which instances to select by approximating the Temporal Weighting Function (TWF) using a mixture of two Gaussians. They applied TWF to every training document. However, it is often the case that terms with informative for a specific time period and informative across the full temporal range of training documents are both included in the training data that affects overall performance of text categorization as these terms are equally weighted in their approach. Moreover, their method needs tagged training data across full temporal range of training documents to create TWF.

There are three novel aspects in our method. Firstly, we propose a method for text categorization that minimizes the impact of temporal effects in both feature selection and learning techniques. Secondly, from manual annotation of data perspective, we propose a temporal-based classification method using only a limited number of labeled training data. Finally, from the perspective of robustness, the method is automated, and can be applied easily to a new domain, or different languages, given sufficient unlabeled documents.

3 SYSTEM DESIGN

The method consists of three steps: (1) Collection of documents by Latent Dirichlet Allocation (LDA), (2) Temporal-based feature selection (TbFS), and (3) Document categorization by temporal-based transfer learning (TbTL).

3.1 Collection of Documents by LDA

The selection of temporal independent/dependent terms is done using documents with categories. However, manual annotation of categories are very expensive and time-consuming. Therefore, we used a topic model and classified unlabeled documents into categories. Topic models such as probabilistic latent semantic indexing (Hofmann, 1999) and LDA (Blei et al., 2003) are based on the idea that documents are mixtures of topics, where each topic is captured by a distribution over words. The topic probabilities provide an explicit low-dimensional representation of a document. They have been successfully used in many tasks such as text modeling and collaborative filtering (Li et al., 2013). We classified documents into categories using LDA. The generative process for LDA can be described as follows:

1. For each topic $k = 1, \cdots, K$, generate ϕ_k, multinomial distribution of terms specific to the topic k from a Dirichlet distribution with parameter β;
2. For each document $d = 1, \cdots, D$, generate θ_d, multinomial distribution of topics specific to the document d from a Dirichlet distribution with parameter α;
3. For each term $n = 1, \cdots, N_d$ in document d;
   (a) Generate a topic z_{dn} of the n^{th} term in the document d from the multinomial distribution θ_d
   (b) Generate a term w_{dn}, the term associated with the n^{th} term in document d from multinomial ϕ_{zdn}

Like much previous work on LDA, we used Gibbs sampling to estimate ϕ and θ. The sampling probability for topic z_i in document d is given by:

$$P(z_i \mid z_{\setminus i}, W) = \frac{(n^{v}_{\setminus i,j} + \beta)(n^{d}_{\setminus i,j} + \alpha)}{(n^{\cdot}_{\setminus i,j} + W\beta)(n^{d}_{\setminus i,\cdot} + T\alpha)}. \quad (1)$$

$z_{\setminus i}$ refers to a topic set Z, not including the current assignment z_i. $n^{v}_{\setminus i,j}$ is the frequency of term v in topic j that does not include the current assignment z_i, and $n^{\cdot}_{\setminus i,j}$ indicates a summation over that dimension. W refers to a set of documents, and T denotes the total number of unique topics. After a sufficient number of sampling iterations, the approximated posterior can be used to estimate ϕ and θ by examining the frequencies of term assignments to topics and topic occurrences in documents. The approximated probability of topic k in the document d, $\hat{\theta}^k_d$, and the assignments term w to topic k, $\hat{\phi}^w_k$ are given by:

$$\hat{\theta}^k_d = \frac{N_{dk} + \alpha}{N_d + \alpha K}. \quad (2)$$

$$\hat{\phi}^w_k = \frac{N_{kw} + \beta}{N_k + \beta V}. \quad (3)$$

For each year, we applied LDA to a set of documents where a set consists of a small number of labeled documents and a large number of unlabeled documents. We need to estimate two parameters for the results obtained by LDA, *i.e.* one is the number of topics/classes k, and another is the number of documents d for each topic/class. We note that the result can be regarded as a clustering result: each element of the clusters is a document assigned to a category or a document without a category information. We estimated the numbers of topics and documents using Entropy measure given by:

$$E = -\frac{1}{\log k}\sum_{j}\frac{N_j}{N}\sum_{i}P(A_i,C_j)\log P(A_i,C_j). \quad (4)$$

k refers to the number of clusters. $P(A_i,C_j)$ is a probability that the elements of the cluster C_j assigned to the correct class A_i. N denotes the total number of elements and N_j shows the total number of elements assigned to the cluster C_j. The value of E ranges from 0 to 1, and the smaller value of E indicates better result. We chose the parameters k and d whose value of E is smallest. For each cluster, count the numbers for each category, and assigned the maximum number of category to each document in the cluster. If there are more than two categories with the maximum numbers, we assigned all of these categories to each document in the cluster.

3.2 Temporal-based Feature Selection

The second step is to select a set of independent/dependent terms from the training data obtained by the first step, collection of documents by LDA. The selection is based on the use of feature selection technique. We tested different feature selection techniques, χ^2 statistics, mutual information, and information gain (Yang and Pedersen, 1997; Forman, 2003). In this paper, we report only χ^2 statistics that optimized global F-score in classification. χ^2 is given by:

$$\chi^2(t,C) = \frac{n\times(ad-bc)^2}{(a+c)\times(b+d)\times(a+b)\times(c+d)}. \quad (5)$$

Using the two-way contingency table of a term t and a category C, a is the number of documents of C containing the term t, b is the number of documents of other class (not C) containing t, c is the number of documents of C not containing the term t, and d is the number of documents of other class not containing t. n is the total number of documents.

We applied χ^2 statistics in two ways. The first way is to extract independent terms that are salient across the full temporal range of training documents. For each category C_i $(1 < i \leq s)$, where s is the number of categories, we collected all documents with the same category across the full temporal range, and created a set. The number of sets equals to the number of categories, s. The second way is to extract dependent terms that are salient for a specific time period. It is applied to the sets of documents with different years in the same category. For a specific category C_i, we collected all documents within the same year, and created a set. Thus, the number of sets equals to the number of different years in the training documents. We selected terms whose χ^2 value is larger than a certain threshold value and regarded these terms as independent/dependent terms.

3.3 Document Categorization

So far, we made use of the maximum amount of tagged old data in feature selection. The final step is document categorization by TbTL. We trained the model and classified documents based on TrAdaBoost (Dai et al., 2007). TrAdaBoost extends AdaBoost (Freund and Schapire, 1997) which aims to boost the accuracy of a weak learner by adjusting the weights of training instances and learn a classifier accordingly. TrAdaBoost uses two types of training data. One is so-called *same-distribution* training data that has the same distribution as the test data. In general, the quantity of these data is often limited. In contrast, another data so-called *diff-distribution* training data whose distribution may differ from the test data is abundant. The TrAdaBoost aims at utilizing the diff-distribution training data to make up the deficit of a small amount of the same-distribution to construct a high-quality classification model for the test data. TrAdaBoost is the same behavior as boosting for same-distribution training data. The difference is that for diff-distribution training instances, when they are wrongly predicted, we assume that these instances do not contribute to the accurate test data classification, and the weights of these instances decrease in order to weaken their impacts. Dai *et al.* applied TrAdaBoost to three text data, 20 Newsgroups, SRZZ, Reuters-21578 which have hierarchical structures. They split the data to generate diff-distribution and same-distribution sets which contain data in different subcategories. Our temporal-based transfer learning, TbTL is based on the TrAdaboost. The difference between TbTL and TrAdaBoost presented by Dai *et al.* is that the initialization step and output the final hypothesis. The initialization step is to remove outliers. The outliers (training instances) are often included in the diff-distribution data itself, especially if

Table 1: Categorization Results (Mainichi data).

Cat	SVM/wo	SVM/w	bSVM/wo	bSVM/w	TrAdaB/wo	TrAdaB/w	TbTL/wo	TbTL/w
International	0.543∗	0.582∗	0.546∗	0.682∗	0.667∗	0.682∗	0.675∗	0.693
Economy	0.564∗	0.594∗	0.699∗	0.702∗	0.665∗	0.702∗	0.672∗	0.712
Home	0.432∗	0.502∗	0.449∗	0.692∗	0.660∗	0.703∗	0.664∗	0.720
Culture	0.082∗	0.102∗	0.158∗	0.301∗	0.459∗	0.493	0.402∗	0.482
Reading	0.468∗	0.489∗	0.563∗	0.571∗	0.662∗	0.697	0.530∗	0.682
Arts	0.353∗	0.372∗	0.387∗	0.652∗	0.656∗	0.663∗	0.664∗	0.693
Sports	0.773∗	0.782∗	0.792∗	**0.802**	0.657∗	**0.730**	0.675∗	0.810
Local news	0.623∗	0.644∗	0.643∗	**0.702**	0.660∗	**0.700**	0.667∗	0.710
Macro Avg.	0.480∗	0.508∗	0.530∗	0.638∗	0.636∗	0.671∗	0.619∗	0.688

∗ denotes that TbTL/w is statistical significance t-test compared with the ∗ marked method, P-value ≤ 0.05

there are a large amount of diff-distribution data. As a result, they affect overall performance of classification. We removed these outliers in the initialization step. The second difference is the output the final hypothesis. We empirically tested Output by both of the TrAdaBoost proposed by Dai *et al.* (Dai et al., 2007) and AdaBoost (Freund and Schapire, 1997), and choose AdaBoost's Output *i.e.* a hypothesis h_t is created at each round by linearly combining the weak hypotheses constructed so far $h_1, \cdots, h_N$ with weights $\beta_1, \cdots, \beta_N$ as it was better to the result obtained by TrAdaBoost, *i.e.* the hypothesis h_t from the $\lceil N/2 \rceil^{th}$ iteration to the N^{th} is voted in the experiments. The temporal-based transfer learning, TbTL based on TrAdaBoost is illustrated in Figure 1.

```
Input {
  The diff-distribution data $Tr_d$, the same-distribution
    data $Tr_s$, and the maximum number of iterations $N$.
}
Output {
  $h_f(x) = \sum_{t=1}^{N} \beta_t h_t(x_i)$.
}
Initialization {
  1. $\mathbf{w^1} = 1/n$.
  2. Train a weak learner on the training set $Tr_d$, and create
     weak hypothesis $h_0$: $X \rightarrow Y$
  3. Classify $Tr_d$ by $h_0$
  4. Create a new diff-distribution training data set
     $Tr_{d\_new}$ where each element $x_i$ satisfies
     $\sum_{i=1}^{n} |h_0(x_i) - c(x_i)| = 0$.
  5. $\mathbf{w^1} = 1/(n' + m)$.
   // $n'$ refers to the number of documents in $Tr_{d\_new}$.
}
TrAdaBoost {
For $t = 1, \cdots, N$
    1. Set $\mathbf{P^t} = \mathbf{w^t} / (\sum_{i=1}^{n'+m} w_i^t)$.
    2. Train a weak learner on the combined training set
       $Tr_{d\_new}$ and $Tr_s$ with the distribution $\mathbf{P^t}$, and create
       weak hypothesis $h_t$: $X \rightarrow Y$
    3. Calculate the error of $h_t$ on $Tr_s$:
       $\varepsilon_t = \sum_{i=n'+1}^{n'+m} \frac{w_i^t \cdot |h_t(x_i) - c(x_i)|}{\sum_{i=n'+1}^{n'+m} w_i^t}$.
    4. Set $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$ and $\beta = 1/(1 + \sqrt{2 \ln n'/N})$.
    5. Update the new weight vector:
       $w_i^{t+1} = \begin{cases} w_i^t \beta^{|h_t(x_i) - c(x_i)|}, & 1 \leq i \leq n' \\ w_i^t \beta_t^{-|h_t(x_i) - c(x_i)|} & n' + 1 \leq i \leq n' + m \end{cases}$
}
```

Figure 1: Flow of the algorithm.

Tr_d shows the diff-distribution training data that $Tr_d = \{(x_i^d, c(x_i^d))\}$, where $x_i^d \in X_d$ (i = 1, $\cdots$, n), and X_d refers to the diff-distribution instance space. Similarly, Tr_s represents the same-distribution training data that $Tr_s = \{(x_i^s, c(x_i^s))\}$, where $x_i^s \in X_s$ (i = 1, $\cdots$, m), and X_s refers to the same-distribution instance space. n and m are the number of documents in T_d and T_s, respectively. $c(x)$ returns a label for the input instance x. The combined training set $T = \{(x_i, c(x_i))\}$ is given by:

$$x_i = \begin{cases} x_i^d & i = 1, \cdots, n \\ x_i^s & i = n+1, \cdots, n+m \end{cases}$$

Steps 2, 3, and 4 of Initialization in Figure 1 are the extraction of outliers that are different term distribution among diff-distribution training data. We removed these training data from the original diff-distribution training data Tr_d, and used the remains Tr_{d_new} as in the input of TrAdaBoost. n' of TrAdaBoost in Figure 1 refers to the number of the remaining diff-distribution training documents.

We used the Support Vector Machines (SVM) as a learner. We represented each training and test document as a vector, each dimension of a vector is an independent/dependent term appeared in the document, and each element of the dimension is a term frequency. We applied the algorithm shown in Figure 1. After several iterations, a learner model is created by linearly combining weak learners, and a test document is classified using a learner.

4 EXPERIMENTS

We evaluated our temporal-based term selection and learning techniques by using the Mainichi Japanese newspaper documents.

4.1 Experimental Setup

We used the Mainichi Japanese newspaper corpus from 1991 to 2012. The corpus consists of 2,883,623 documents organized into 16 categories. We selected 8 categories, “International”, “Economy”, “Home”, “Culture”, “Reading”, “Arts”, “Sports”, and “Local news”, each of which has sufficient number of documents. Table 2 shows statistics of the dataset.

Table 2: The data used in the experiments.

Cat	Docs	Cat	Docs
International	91,882	Reading	17,418
Economy	96,745	Arts	29,645
Home	47,984	Sports	183,216
Culture	20,428	Local news	282,829

All documents were tagged by using a morphological analyzer Chasen (Matsumoto et al., 2000). We used noun words as independent/dependent term selection. The total number of documents assigned to these categories are 770,147. For each category within each year, we divided documents into three folds: 10% of documents are used as labeled training data, 50% of documents are unlabeled training data, and 40% of documents are used to test our classification method. For each year, we classified unlabeled data into categories using labeled data with LDA. We empirically selected values of two parameters, the number of classes k, and documents d, respectively. k is searched in steps of 10 from 10 to 200, and d is searched in steps of 100 from 100 to 500. As a result, for each year, we set k and d to 20, and 700, respectively.

We divided original labeled training data and labeled data obtained by LDA into five folds for each year. The first three folds are used in the TbFS, *i.e.* we calculated χ^2 statistics using the first fold, and the second fold is used as a training data and the third fold is used as a test data to estimate the numbers of independent/dependent terms. The estimation was done by using F-score. As a result of estimation, we used 35,000 independent terms for each of the 8 categories, and 12,000 dependent terms for each of the 8 categories in each year. The last two folds are used to train TbTL. For each category, we used 50 documents as the same-distribution data. When the time difference between training and test data is more than one year, we used the remains as diff-distribution data[2].

We used SVM-light (Joachims, 1998) as a basic learner in the experiments. We compared our method, TbTL with TbFS (TbTL/w) with seven baselines: (1) SVM without TbFS (SVM/wo), (2) SVM with TbFS (SVM/w), (3) biased-SVM (Liu et al., 2003) without TbFS (bSVM/wo), (4) biased-SVM with TbFS (bSVM/w)), (5) TrAdaBoost without TbFS (TrAdaB/wo), (6) TrAdaBoost with TbFS (TrAdaB/w), and (7) TbTL without TbFS (TbTL/wo). The methods without TbFS, *i.e.* (1), (3), (5), and (7), we used all noun words in the documents.

TrAdaBoost refers to the results obtained by original TrAdaBoost presented by Dai *et al.* Biased-SVM is known as the state-of-the-art SVMs method, and often used for comparison (Elkan and Noto, 2008). Similar to the SVM, for biased-SVM, we used the last two folds as a training data, and classified test documents directly, *i.e.* we used closed data. We empirically selected values of two parameters, “*c*” (trade-off between training error and margin) and “*j*”, *i.e.* cost (cost-factor, by which training errors on positive examples) that optimized F-score obtained by classification of test documents. “*c*” is searched in steps of 0.02 from 0.01 to 0.61. Similarly, “*j*” is searched in steps of 5 from 1 to 200. As a result, we set *c* and *j* to 0.03 and 4, respectively. To make comparisons fair, all eight methods including our method are based on linear kernel. Throughout the experiments, the number of iterations is set to 30.

4.1.1 Results

Categorization results for 8 categories (40% of the test documents, *i.e.* 308,058 documents) are shown in Table 1. Each value in Table 1 shows macro-averaged F-score across 22 years. “Macro Avg” in Table 1 refers to macro-averaged F-score across categories. The results obtained by biased-SVM indicate the maximized F-score obtained by varying the parameters, “*c*” and “*j*”. As can be seen clearly from Table 1, the results with “TbTL/w” and “TrAdaB/w” were better to the results obtained by “bSVM/w” except for “Sports” and “Local news” in “TrAdaB/w”, although “bSVM/w” in Table 1 was the result obtained by using the closed data. Moreover, the results obtained by SVM with and without TbFS was the worst result among other methods. These observations show that once the training data drive from a different time period from the test data, the distributions of terms between training and test documents are not identical.

[2] When the creation time period of the training data is the same as the test data, we used only the same-distribution data.

Table 3: Sample results of term selection.

Sports		International	
ind.	dep. (2000)	ind.	dep. (1997)
baseball	Sydney	president	Tupac Amaru
win	Toyota	premier	Lima
game	HP	army	Kinshirou
competition	hung-up	power	residence
championship	Paku	government	Hirose
entry	admission	talk	Huot
tournament	game	election	MRTA
player	Mita	UN	Topac
defeat	Miyawaki	politics	impression
pro	ticket	military	employment
title	ready	nation	earth
finals	Seagirls	democracy	election
league	award	minister	supplement
first game	Gaillard	North Korea	East Europe
Olympic	attackers	chair	bankruptcy

The overall performance with TbFS were better to those without TbFS in all methods. This shows that temporal-based term selection contributes classification performance. Table 3 shows the topmost 15 independent/dependent terms obtained by TbFS. The categories are "Sports" and "International". As we can see from Table 3 that independent terms such as "baseball" and "win" are salient terms of the category "Sports" regardless to a time period. On the other hand, "Miyawaki" listed in the dependent terms. The term often appeared in the documents from 1998 to 2000 because Miyawaki was a snowboard player and he was on his first world championship title in Jan. 1998. Similarly, in the category "International", terms such as "UN" and "North Korea" are listed in the independent terms, as they often appeared in documents regardless of the timeline. In contrast, "Tupac Amaru" and "MRTA" are listed in the dependent terms. It is reasonable because in this year, Tupac Amaru Revolutionary Movement (MRTA) rebels were all killed when Peruvian troops stormed the Japanese ambassador's home where they held 72 hostages for more than four months. These observations support our basic assumption: there are two types of salient terms, *i.e.* terms that are salient for a specific period, and terms that are important regardless of the timeline.

Figures 2 and 3 illustrate F-score with/without TbFS against the temporal difference between training and test data. Both training and test data are the documents from 1991 to 2012. For instance, "5" of the x-axis in Figures 2 and 3 indicate that the test documents are created 5 years later than the training documents. We can see from Figures 2 and 3 that the results with TbFS were better to those without TbFS in all of the methods. Moreover, the result obtained by "TbTL/w" in Figure 2 was the best in all of the temporal distances. There are no significant differences among three methods, "bSVM", "TrAdaB", and "TbTL" when the test and training data are the same time period in both of the Figures 2 and 3. The performance of these methods including "SVM" drops when the period of test data is far from the training data. However, the performance of "TbTL" was still better to those obtained by other methods. This demonstrates that the algorithm which applies temporal-based feature selection and learning is effective for categorization. Figure 4 shows the averaged F-score of categories across full temporal range with TbFS against the number of iterations. Although the curves are not quite smooth, they converge around 25 iterations.

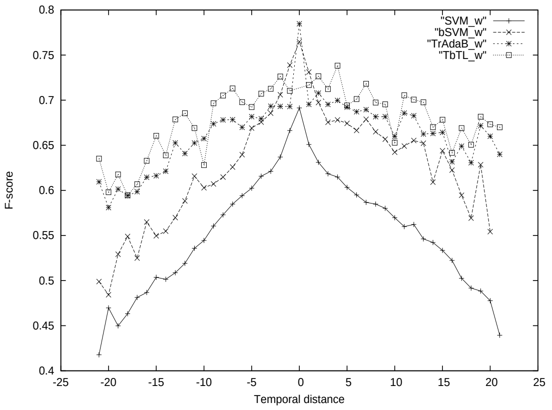

Figure 2: Performance with TbFS against temporal distance.

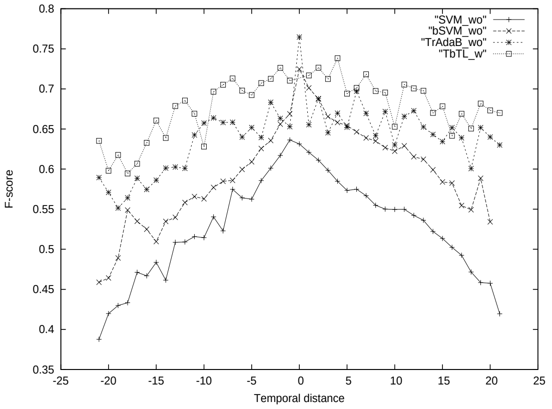

Figure 3: Performance without TbFS against temporal distance.

Finally, we tested how the use of LDA influences the overall performance. Figure 5 illustrates F-score of "TbTL/w" with and without LDA against the temporal difference between training and test data. In "TbTL/w" without LDA, we added 50% (393,759) labeled documents to the original 10% (78,751) la-

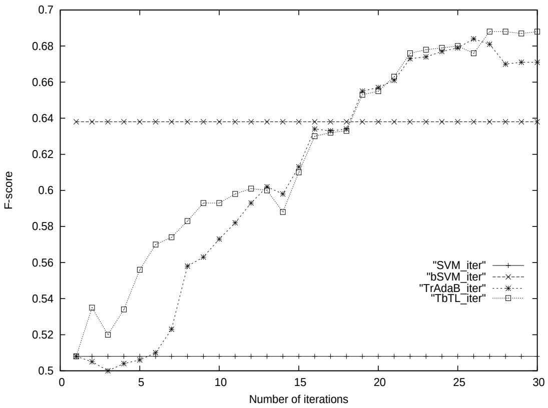

Figure 4: F-score with TbFS against the # of iterations.

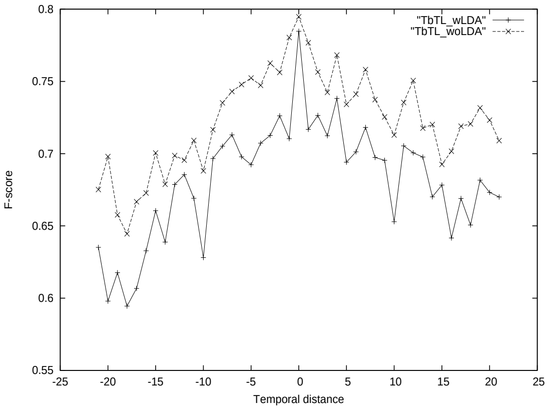

Figure 5: F-score with/without LDA against temporal distance.

beled training documents. As we expected, the results obtained by “TbTL/w” without LDA were better to those with LDA in every temporal distance, and the averaged improvement of F-score across 22 years was 3.5%(0.723-0.688). It is not surprising because in “TbTL/w” without LDA, we used a large number of labelled training documents, 472,510 documents which are very expensive and time-consuming. In contrast, in “TbTL/w” with LDA, we used 78,751 labeled documents across 22 years in all, the average number of documents per year was 3,579 across eight categories.

5 CONCLUSIONS AND FUTURE WORK

We have developed an approach for text categorization concerned with the impact that the variation of the strength of term-category relationship over time. The basic idea is to minimize the impact of temporal effects in both feature selection and learning techniques. The results using Japanese Mainichi Newspaper corpus show that temporal-based feature selection and learning method works well for categorization, especially when the creation time of the test data differs greatly from the training data.

There are a number of interesting directions for future work. We should be able to obtain further advantages in accuracy in independent/dependent term selection by smoothing the term distributions such as organization and person name terms through the use of techniques such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Log-Bilinear Document Model (Maas and Ng, 2010), and word2vec (Mikolov et al., 2013). The quantity of the labeled training documents affects the overall performance. Dai *et al.* attempted to use Transductive Support Vector Machines (Dai et al., 2007; Joachims, 1999). However, they reported that the rate of convergence is slow. This issue needs further investigation. We used LDA to classify unlabeled documents into categories. There are number of other topic models such as continuous time dynamic topic model (Wang et al., 2008) and a biterm topic model (Yan et al., 2013). It is worth trying to test these methods for further improvement.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their comments on the earlier version of this paper. This work was supported by the Grant-in-aid for the Japan Society for the Promotion of Science (JSPS), No. 25330255.

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Machine Learning,* 3:993–1022.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain Adaptation with Structural Correspondence Learning. In *Proc. of the Conference on Empirical Methods in Natural Language Processing,* pp. 120-128.

Dai, W., Yang, Q., Xue, G., and Yu, Y. (2007). Boosting for Transfer Learning. In *Proc. of the 24th International Conference on Machine Learning,* pp. 193-200.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Hashman, R. (1990). Indexing by Latent Semantic Analysis. *American Society for Information Science*, 41(6):391–407.

Dumais, S. and Chen, H. (2000). Hierarchical Classification of Web Contents. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 256-263.

Elkan, C. and Noto, K. (2008). Learning Classifiers from Only Positive and Unlabeled Data. In *Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* pp. 213-220.

Folino, G., Pizzuti, C., and Spezzano, G. (2007). An Adaptive Distributed Ensemble Approach to Mine Concept-drifting Data Streams. In *Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence,* pp. 183-188.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Machine Learning Research*, 3:1289–1305.

Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In *Proc. of the 28th International Conference on Machine Learning,* pp. 97-110.

Gopal, S. and Yang, Y. (2010). Multilabel Classification with Meta-level Features. In *Proc. of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 315-322.

Hassan, S., Mihalcea, R., and Nanea, C. (2007). Random-Walk Term Weighting for Improved Text Classification. In *Proc. of the IEEE International Conference on Semantic Computing,* pp. 242-249.

He, D. and Parker, D. S. (2010). Topic Dynamics: An Alternative Model of Bursts in Streams of Topics. In *Proc. of the 16th ACM SIGKDD Conference on Knowledge discovery and Data Mining,* pp. 443-452.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 35-44.

III, H. D. (2007). Frustratingly Easy Domain Adaptation. In *Proc. of the 45th Annual Meeting of the Association of computational Linguistics,* pp. 256-263.

Joachims, T. (1998). SVM Light Support Vector Machine. In *Dept. of Computer Science Cornell University*.

Joachims, T. (1999). Transductive Inference for Text Classification using Support Vector Machines. In *Proc. of 16th International Conference on Machine Learning,* pp. 200-209.

Kerner, Y. H., Mughaz, D., Beck, H., and Yehudai, E. (2008). Words as Classifiers of Documents according to Their Historical Period and the Ethnic Origin of Their Authors. *Cymernetics and Systems*, 39(3):213–228.

Kleinberg, M. (2002). Bursty and Hierarchical Structure in Streams. In *Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 91-101.

Klinkenberg, R. and Joachims, T. (2000). Detecting Concept Drift with Support Vector Machines. In *Proc. of the 17th International Conference on Machine Learning,* pp. 487-494.

Lazarescu, M. M., Venkatesh, S., and Bui, H. H. (2004). Using Multiple Windows to Track Concept Drift. *Intelligent Data Analysis*, 8(1):29–59.

Lewis, D. D. and Ringuette, M. (1994). Comparison of Two Learning Algorithms for Text Categorization. In *Proc. of the Third Annual Symposium on Document Analysis and Information Retrieval,* pp. 81-93.

Li, Y., Yang, M., and Zhang, Z. (2013). Scientific Articles Recommendation. In *Proc. of the ACM International Conference on Information and Knowledge Management CIKM 2013,* pp. 1147-1156.

Liu, B., dai, Y., Li, X., Lee, W. S., and Yu, P. S. (2003). Building Text Classifiers using Positive and Unlabeled Examples. In *Proc. of the ICDM'03,* pp. 179-188.

Maas, A. L. and Ng, A. Y. (2010). A probabilistic Model for Semantic Word Vectors. *NIPS*, 10.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, Y., Takaoka, K., and Asahara, M. (2000). *Japanese Morphological Analysis System Chasen Version 2.2.1*. In Naist Technical Report.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proc. of the International Conference on Learning Representations Workshop*.

Mourao, F., Rocha, L., Araujo, R., Couto, T., Goncalves, M., and Jr., W. M. (2008). Understanding Temporal Aspects in Document Classification. In *Proc. of the 1st ACM International Conference on Web Search and Data Mining,* pp. 159-169.

Murphy, J. (1999). *Technical Analysis of the Financial Markets*. Prentice Hall.

Raina, R., Ng, A. Y., and Koller, D. (2006). Constructing Informative Priors using Transfer Learning. In *Proc. of the 23rd International Conference on Machine Learning,* pp. 713-720.

Salles, T., Rocha, L., and Pappa, G. L. (2010). 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.

Siao, M. and Guo, Y. (2013). Domain Adaptation for Sequence Labeling Tasks with a Probabilistic Language Adaptation Model. In *Proc. of the 30th International Conference on Machine Learning,* pp. 293-301.

Song, M., Heo, G. E., and Kim, S. Y. (2014). Analyzing topic evolution in bioinformatics: Investigation of dynamics of the field with conference data in dblp. *Scientometrics*, 101(1):397–428.

Sparinnapakorn, K. and Kubat, M. (2007). Combining Subclassifiers in Text Categorization: A DST-based Solution and a Case Study. In *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 210-219.

Wang, C., Blei, D., and Heckerman, D. (2008). Continuous Time Dynamic Topic Models. In *Proc. of the 24th Conference on Uncertainty in Artificial Intelligence,* pp. 579-586.

Xue, G. R., Dai, W., Yang, Q., and Yu, Y. (2008). Topic-bridged PLSA for Cross-Domain Text Classification.

In *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 627-634.

Yan, X., Guo, J., Lan, Y., and X.Cheng (2013). A Biterm Topic Model for Short Texts. In *Proc. of the 22nd International Conference on World Wide Web,* pp. 1445-1456.

Yang, Y. and Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In *Proc. of the 14th International Conference on Machine Learning,* pp. 412-420.