The Effect of Temporal-based Term Selection for Text Classification

Fumiyo Fukumoto[1], **Shogo Ushiyama**[2], **Yoshimi Suzuki**[1] and **Suguru Matsuyoshi**[1]
[1]Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi
[2]Faculty of Engineering, University of Yamanashi
Kofu, 400-8511, JAPAN
{fukumoto, t08kg006, ysuzuki, sugurum}@yamanashi.ac.jp

Abstract

This paper addresses the text classification problem that training data may derive from a different time period from the test data. We present a method of temporal-based term selection for timeline adaptation. We selected two types of informative terms according to corpus statistics. One is temporal independent terms that are salient regardless of the timeline. Another is temporal dependent terms which are important for a specific period of time. For temporal dependent terms extracted from the training documents, we applied weighting function that weights terms according to the temporal distance between training and test data in the process of training classifiers. The results using Mainichi Japanese newspaper documents showed improvement over the three baselines.

1 Introduction

Text classification supports and improves several tasks such as automated 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 techniques have been applied to text classification (Xue et al., 2008; Gopal and Yang, 2010). The common approach is the use of term selection. Each document is represented using a vector of selected terms (Yang and Pedersen, 1997; Hassan et al., 2007). Then, they used training documents with category label to train classifiers. Once category models are trained, each document of the test data is classified by using these models. Terms in the documents may be considered more important to build the classification model according to the timelines, while the majority of supervised classification methods consider that each term provides equally information regardless to a period. For instance, as shown in Figure 1, the term “earthquake” appeared more frequently in the category “Science” than “International” early in 1995. However, it appeared frequently in the category “International” than “Science” since Sumatra earthquake occurred just off the southern coast of Sumatra, Indonesia in 2005. Similarly, the term “Alcindo” frequently appeared in the documents tagged “Sports” in 1994, since Alcindo is a Brazilian soccer player and he was one of the most loved players in 1994. The term did not appear more frequently in the “Sports” category since he retired in 1997. These observations show that salient terms in the training data, are not salient in the test data when training data may derive from a different time period from the test data.

In this paper, we present a method for text classification concerned with the impact that the variation of the strength of term-class relationship over time. We selected two types of informative terms according to corpus statistics. One is temporal independent terms that are salient regardless of the timeline. Another is temporal dependent terms which are salient for a specific period of time. For temporal dependent terms extracted from the training documents, we applied weighting function that weights terms according to the temporal distance between training and test data in the process of training classifiers.

Our weighting function is based on an algorithm called temporally-aware algorithm that used a Temporal Weighting Function (TWF) developed by Salles *et al.* (Salles et al., 2010). The method incorporates temporal models to document classifiers. The weights assigned to each document depend on the notion of a temporal distance, defined as the difference between the time of creation of a training example and a reference time point, *i.e.*,

Fumiyo Fukumoto, Shougo Ushiyama, Yoshimi Suzuki and Suguru Matsuyoshi. 2014. The Effect of Temporal-based Term Selection for Text Classification. In *Proceedings of Australasian Language Technology Association Workshop*, pages 88−96.

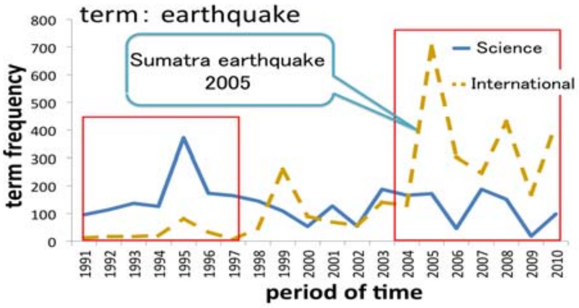

Figure 1: "earthquake" appeared in "Science" and "International" categories

temporal weighting weights training instances according to the temporal distance between training and test instances. The difference is that we applied the function to only dependent terms while a method of Salles weights all terms in the training documents. Because as illustrated in Figure 1, "earthquake" that are salient for a specific period of time and terms such as "Science" which are important regardless of the timeline in "Science" domain are both included in the training documents. These terms appearing in the training documents are equally weighted, which affect classification accuracy.

The remainder of the paper is organized as follows: Section 2 describes related work. Section 3 briefly reviews temporally-aware algorithm. Section 4 presents our framework. Finally, we report experiments and conclude our discussion with some directions for further work.

2 Related Work

The analysis of temporal aspects for text classification is a practical problem attracting more and more attention. Mourao *et al.* have shown evidence that time is an important factor in text classification (Mourao et al., 2008). More specifically, they selected training documents that are closer in time to the test document. They reported that the method has attained at 89.8% accuracy for ACM, and 87.6% for Medline. Cohen *et al.* attempted to extract context including phrases that is exploited towards better classification models (Cohen and Singer, 1999). Kim *et al.* focused on Web documents and presented a classification method using the knowledge acquisition method, Multiple Classification Ripple Down Rules (MCRDR). It enables domain users to elicit their domain knowledge incrementally and revise their knowledge base. They may then reclassify documents according to context changes (Kim et al., 2004). These techniques can be classified into adaptive document classification (Yang and Lin, 1999; Dumais and Chen, 2000; Liu and Lu, 2002; Rocha et al., 2008) where temporal aspects are considered to classification.

Several authors have attempted to capture concept or topic drift dealing with temporal effects in classification (Kelly et al., 1999; Lazarescu et al., 2004; Folino et al., 2007; Ross et al., 2012). The earliest known approach is the work of (Klinkenberg and Joachims, 2000). They attempted to handle concept changes with SVM. They used $\xi\alpha$-estimates to select the window size so that the estimated generalization error on new examples is minimized. The result which was tested on the TREC shows that the algorithm achieves a low error rate and selects appropriate window sizes. Scholz *et al.* proposed a method called knowledge-based sampling strategy (KBS) to train a classifier ensemble from data streams. They used two types of data sets, 2,608 documents of the data set of the TREC, and the satellite image dataset from the UCI library to evaluate their method. They showed that the algorithm outperformed leaning algorithms without considering concept drift (Scholz and Klinkenberg, 2007). He *et al.* attempted 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 their method using MeSH terms and reported that the model works well for tracking topic bursts.

As mentioned above, several efforts have been made to automatically identify context changes, topic drift or topic bursts. Most of these focused just on identifying the increase of a new context, and not relating these contexts to their chronological time. In contrast, we propose a method that minimizes temporal effects to achieve high classification accuracy. In this context, Salles *et al.* proposed 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. They addressed the drawbacks of which instances to select by approximating the Temporal Weighting Function (TWF) us-

Table 1: Temporal distances against terms

	t_1	t_2	$\cdots$	t_k	D_δ
δ_1	f_{11}	f_{12}	$\cdots$	f_{1k}	$\sum_{i=1}^{k} f_{1i}$
δ_2	f_{21}	f_{22}	$\cdots$	f_{2k}	$\sum_{i=1}^{k} f_{2i}$
$\vdots$					
δ_n	f_{n1}	f_{n2}	$\cdots$	f_{nk}	$\sum_{i=1}^{k} f_{ni}$

ing a mixture of two Gaussians. They applied TWF to every training document. However, it is often the case that terms with salient for a specific period of time and important terms regardless of the timeline are both included in the training documents. We focus on the issue, and present an algorithm which weights only to the salient terms in a specific period of time.

3 Temporal Weighting Function

In this section, we briefly review Temporal Weighting Function (TWF) proposed by Salles *et al.* (Salles et al., 2010). TWF is based on the temporal distance between training and test documents creation times (Salles et al., 2010). Given a test document to be classified, the TWF sets higher weights to training documents that are more similar to the test document. The weights refer to the strength of term-class relationships. It is defined as $dominance(t, c) = \frac{N_{tc}}{\sum_{c'} N_{tc'}}$ where N_{tc} refers to the number of documents in class c that contain term t. When the dominance $dominance(t, c)$ is larger than a certain threshold value α[1], the term is judged to have a high degree of exclusivity with some class.

We note that TWF sets higher weights to training documents that temporally close to the test document. Let $S^{'}_{t}$ = $\{\delta \leftarrow p_n - p_r \mid \forall r p_n \in S_{t,r}\}$ be a set of temporal distances that occur on the stability periods of term t. Here, p_n be the time of creation concerning to a training document. Stability periods of term t, referred to as $S_{t,r}$ is the largest continuous period of time, starting from the reference time point p_r in which the test document was created and growing both to the past and the future. For instance, if $S_{t,r}$ is $\{$1999, 2000,2001$\}$, and p_r = 2000, then $S^{'}_{t}$ = $\{$-1,0,1$\}$.

Finally, the function is determined considering the stability period of each term as a random variable where the occurrence of each possible temporal distance in its stability period is an event. The frequencies of the temporal distances δ_1 to δ_n for terms t_1 to t_k are shown in Table 1. The random variable D_δ related to the occurrences of δ, which represents the distribution of each δ_i over all terms t, is lognormally distributed if InD_δ is normally distributed. Here, InD_δ refers to lognormal distribution D_δ where D_δ stands for the distribution of temporal distance δ_i for the term t_i over all terms t. A 3-parameter Gaussian function, $F = a_i e^{-\frac{(x-b_i)^2}{2c_i^2}}$ is used to estimate the relationship between temporal distance and temporal weight, where the parameter a_i is the height of the curve's peak, b_i is the position of the center of the peak, and c_i controls the width of the curve. These parameters are estimated by using a Maximum Likelihood method.

4 Framework of the System

The method for temporal-based classification consists of three steps: selection of temporal independent/dependent terms, temporal weighting for dependent terms, and text classification.

4.1 Independent and dependent term selection

The first step is to select a set of independent/dependent terms from the training data. 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:

$$\chi^2(t, C) \quad = \quad \frac{n \times (ad - bc)^2}{(a+c) \times (b+d) \times (a+b) \times (c+d)} \tag{1}$$

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.

Independent terms are salient across the full temporal range of training documents. For each category C_i ($1 < i \leq n$), where n is the number of categories, we collected all documents with

[1] We empirically set alpha = 50% in the experiment.

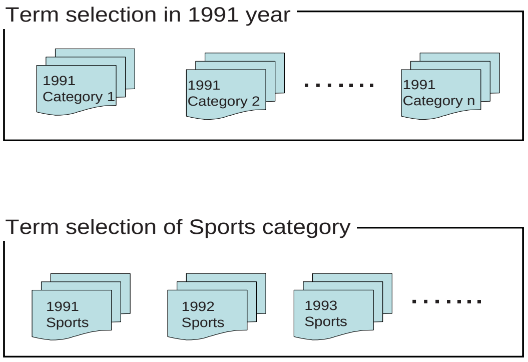

Figure 2: Term selection per year versus category

the same category across the full temporal range, and created a set. The number of sets equals to the number of categories. In contrast, dependent terms refer to a term that is salient for a specific period of time.

As illustrated in Figure 2, we selected dependent terms by using two methods: selection per year, and category. The former is applied to the sets of documents with different categories in the same year as illustrated in the top of Figure 2. For each category in a specific year y_j (y_j in Figure 2 refers to 1991), we collected all documents tagged in the category C_i within the year y_j, and created a set. The number of sets equals to the number of categories in the training documents. In contrast, term selection per category is applied to the sets of documents with different years in the same category shown in the bottom of Figure 2. For a specific category C_i (C_i refers to "Sports" in Figure 2), we collected all documents in the same year, and created a set. Thus, the number of sets equals to the number of different years in the training documents.

4.2 Temporal weighting

We applied χ^2 statistics and selected terms whose χ^2 value is larger than a certain threshold value. The procedure for temporal weighting of the selected term t in the training document D is shown in Figure 3. For each term t in the training document D, if t is included in a set obtained by dependent term selection, t is weighted by TWF, as the term t is salient for a specific year δ. As shown in 4 of Figure 3, if t occurs in several years, we pick up the latest year δ' and t is weighted by TWF(δ'). Because δ' is close to the year that the test document was created, and it can be considered to be reliable for accurate classification. X_{dt} in Figure 3 refers to a set of terms obtained by term selection per year (Year), or term selection per category (Cat).

```
For each term $t \in D$ {
1. If $t$ is included in the set $X_{idt}$, TWF is not
   applied to $t$.
  // $X_{idt}$ is a set of terms obtained by
     independent term selection.
2. Else if $t$ is included in the set $X_{dt}$, {
  // $X_{dt}$ is a set of terms obtained by
     dependent term selection.
3.   If $t$ appears only in a specific year $\delta$, $t$
     is weighted by TWF($\delta$).
4.   Else if $t$ occurs in several years, pick
     the latest year $\delta'$, $t$ is weighted by
     TWF($\delta'$).
   }
}
```

Figure 3: Temporal weighting procedure

4.3 Classification based on kNN

Similar to Salles's experiments (Salles et al., 2010), we tested Rocchio, kNN and NB with the TWF. As a result, we used kNN in the experiment because the result obtained by kNN was the best among them. Each training document is represented using a vector of selected independent/dependent terms. Given a test document, the system finds the k nearest neighbors among the training documents, and uses the categories of the k neighbors to weight the category candidates. The similarity score between training, and test documents collected from 1994 is illustrated in Figure 4.

The graph on the right hand side shows TWF described in Section 3. $Sim(d, d')$ indicates the similarity between training document d and the test document d'. As shown in Figure 4, we used the cosine value of two vectors to measure the similarity between the training and test documents. $f(t)$ refers to the frequency of a term t in the training/test document. t_1 in Figure 4 refers to a term that is important regardless of the timeline. In contrast, t_5 and t_7 are salient terms at a specific year, *i.e.*, 1991 and 1993. These terms are weighted by TWF, *i.e.*, the weight of t_5 is TWF(3) = TWF(1994-1991), and t_7 is TWF(1) = TWF(1994-1993). By sorting the score of candidate categories, a ranked list is obtained for the test

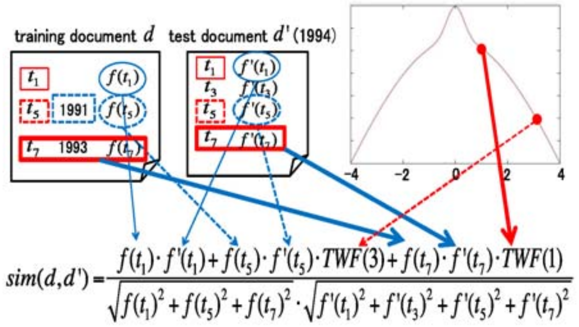
training document d
test document d' (1994)
t_1
$f(t_1)$
t_5
1991
$f(t_5)$
t_7
1993
$f(t_7)$
t_1
$f'(t_1)$
t_3
$f'(t_3)$
t_5
$f'(t_5)$
t_7
$f'(t_7)$
−4
−2
0
2
4
$sim(d,d') = \frac{f(t_1)\cdot f'(t_1)+f(t_5)\cdot f'(t_5)\cdot TWF(3)+f(t_7)\cdot f'(t_7)\cdot TWF(1)}{\sqrt{f(t_1)^2+f(t_5)^2+f(t_7)^2}\cdot\sqrt{f'(t_1)^2+f'(t_3)^2+f'(t_5)^2+f'(t_7)^2}}$

Figure 4: The similarity between training and test documents

document. The category with the highest score of the vote is assigned to the test document.

5 Experiments

We had an experiment to evaluate our method. We collected Mainichi Japanese newspaper from 1991 to 2010 and used them in the experiments[2]. Mainichi newspaper documents are classified into sixteen categories. Of these, we used six categories, "Sports", "Home", "Science", "Economy", "Arts", and "International", each of which has more than 250 documents for each year. All Japanese documents were tagged by using a morphological analyzer Chasen (Matsumoto et al., 2000). We used nouns as independent/dependent term selection.

We divided all the documents into two sets: one is to estimate the number of selected terms weighted by χ^2 statistics, 3 parameters Gaussian function, and the number of k in kNN. Another is a test data. We used the estimated parameters to classify test documents. Table 2 shows the size of data used in the experiments. "Doc" refers to the number of documents per category. As shown in Table 2, we used three types of test data to examine the effect of the method against the difference of period between the training and test data. As a result of parameter estimation, we used the number of 10,000 terms as independent terms and 3,000 for dependent terms. The estimated parameters used in a Gaussian function are shown in Table 3. The number of k in kNN was set to 12.

Table 2: Data used in the experiments

Parameter estimation				
Period	Training		Test	
	Doc	Total	Doc	Total
1991 - 2000	80	4,800	50	3,000
2001	–	–	500	3,000
2010	–	–	500	3,000

Training and Test				
Period	Training		Test	
	Doc	Total	Doc	Total
1991 - 2000	120	7,200	50	3,000
2001	–	–	500	3,000
2010	–	–	500	3,000

Table 3: Estimated parameters

Param.	Value
a_1	0.969
b_1	6.104×10^{-9}
c_1	7.320
a_2	0.031
b_2	-3.451×10^{-7}
c_2	0.506

We evaluated text classification performance by F-score. To examine the effect of dependent term selection, we set X_{dt} in Figure 3 to three types of terms, *i.e.*, Year, Cat, and Year ∪ Cat, and compared these results. Table 4 shows the results obtained by using three types of terms. "Cat" and "Year" refer to the results obtained by term selection per category, and year, respectively. "Cat ∪ Year" refers to the results obtained by both selection methods. "Macro Avg." in Table 4 indicates macro-averaged F-score. "$*$" in Table 4 shows that "Cat" shows statistical significance t-test compared with the $*$ marked method.

As shown in Table 4, there is no significant difference among three selection methods, especially when the test and training documents are the same time period, *i.e.*, 1991 - 2000. When the test data is derived from 2001 and 2010, the macro-averaged F-score obtained by "Cat" is statistically significant compared with "Year" in some categories. These observations indicate that term selection per category is the best among other methods. Then, we used term selection per category as a dependent term selection.

We compared our method, temporal-based term selection(TTS) with three baselines: (1) SVM, (2) kNN, and (3) a method developed by Salles *et al.*

[2] http://ndk.co.jp/english/index.html

Table 4: Classification Results

1991 - 2000 Test Data			
Category	Cat	Year	Cat $\cup$ Year
Arts	0.813	0.819	0.814
International	0.836	0.833	0.835
Economy	0.802	0.802	0.801
Home	0.747	0.751	0.751
Science	0.807	0.806	0.809
Sports	0.920	0.921	0.920
Macro Avg.	0.821	0.822	0.822

2001 Test Data			
Category	Cat	Year	Cat $\cup$ Year
Arts	0.799	0.791*	0.800
International	0.801	0.801	0.803
Economy	0.792	0.789	0.791
Home	0.745	0.740	0.744
Science	0.714	0.713	0.715
Sports	0.897	0.892*	0.898
Macro Avg.	0.791	0.788*	0.792

2010 Test Data			
Category	Cat	Year	Cat $\cup$ Year
Arts	0.330	0.323*	0.322*
International	0.718	0.714	0.718
Economy	0.694	0.698	0.695
Home	0.494	0.501	0.490
Science	0.495	0.496	0.496
Sports	0.862	0.865	0.863
Macro Avg.	0.598	0.600	0.597

* denotes statistical significance t-test, P-value $\leq$ 0.05

Table 5: Comparative results

1991 - 2000 Test Data				
Category	kNN	Salles	SVM	TTS
Arts	0.785*	0.795*	0.801	0.813
International	0.811*	0.810*	0.837	0.836
Economy	0.796	0.799	0.800	0.802
Home	0.715*	0.721*	0.740	0.747
Science	0.803	0.807	0.809	0.807
Sports	0.885*	0.890*	0.892*	0.920
Macro Avg.	0.799*	0.804*	0.812	0.821

2001 Test Data				
Category	kNN	Salles	SVM	TTS
Arts	0.765*	0.764*	0.780*	0.799
International	0.780*	0.783*	0.802	0.801
Economy	0.797	0.805	0.809	0.792
Home	0.717*	0.722*	0.728*	0.745
Science	0.720	0.720	0.723	0.714
Sports	0.867*	0.862*	0.870*	0.897
Macro Avg.	0.774*	0.776*	0.785	0.791

2010 Test Data				
Category	kNN	Salles	SVM	TTS
Arts	0.339	0.310*	0.340	0.330
International	0.688*	0.685*	0.687*	0.718
Economy	0.688	0.676*	0.689	0.694
Home	0.482*	0.477*	0.483*	0.494
Science	0.490	0.478	0.492	0.494
Sports	0.851*	0.850*	0.851*	0.862
Macro Avg.	0.589*	0.579*	0.590*	0.598

* denotes statistical significance t-test, P-value $\leq$ 0.05

(Salles et al., 2010), *i.e.*, the method applies TWF to each document. In SVM and kNN, we used the result of a simple χ^2 statistics. We used SVM-Light package for training and testing (Joachims, 1998)[3]. We used linear kernel and set all parameters to their default values. The results are shown in Table 5. "$*$" in Table 5 shows that TTS is statistical significance t-test compared with the $*$ marked methods. For instance, the performance of "Cat" in category "Arts" by using 1991-2000 test data shows significantly better to the results obtained by both kNN and Salles *et al.* methods.

As can be seen from Table 5 that macro-averaged F-score obtained by TTS was better to those obtained by kNN and Salles's methods in all of the three types of test data. When we used 1991 - 2000 and 2001 test data, the performance against the categories except for "Economy" and "Science" obtained by TTS was better to those obtained by kNN and Salles's methods. The performance obtained by TTS was better than Salles's method, especially the test data (2010) was far from the training data (1991 - 2000), as five out of six categories were statistically significant. These observations show that the algorithm which applies TWF to each term is more effective than the method applying TWF to each document in the training data. There is no significant difference between the results obtained by SVM and TTS when the test data is not far from the training data, *i.e.*,

[3] http://svmlight.joachims.org

Table 6: 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**	Eastern Europe
Olympic	attackers	chair	bankruptcy

1991 - 2000 and 2001. However, when we used 2010 test data, the result obtained by TTS is statistically significant compared with SVM. The observation shows that our method is effective when testing on data far from the training data.

Table 6 shows topmost 15 terms obtained by independent and dependent term selection. The dependent term selection is a result obtained by term selection per category. The categories are “Sports” and “International”. As we can see from Table 6 that independent terms such as “baseball” and “win” are salient terms of the category “Sports” regardless to a time period. In contrast, “Miyawaki” listed in the dependent terms, is a snowboard player and he was on his first world championship title in Jan. 1998. The term often appeared in the documents from 1998 to 2000. Similarly, in the category “International”, terms such as “UN” and “North Korea” often appeared in documents regardless of the timeline, while “Tupac Amaru” and “MRTA” frequently appeared in a specific year, 1997. 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.

We recall that the overall performance obtained by four methods including our method drops when we used 2010 test data, while the performance of our method was still better than other methods in

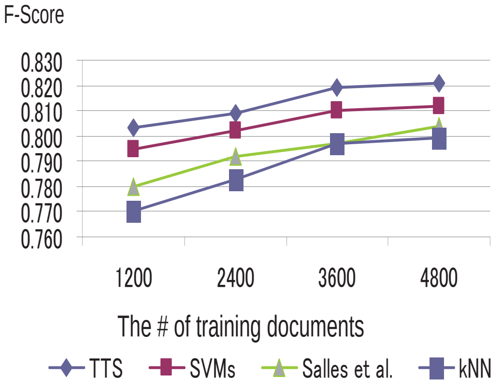

Figure 5: Performance (1991 - 2000 data)

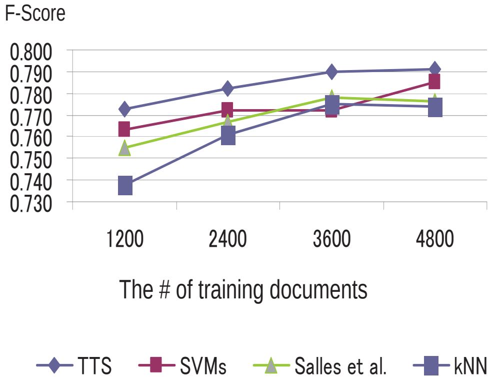

Figure 6: Performance (2001 data)

Table 5. We note that we used surface information, *i.e.*, noun words in documents as a feature of a vector. Therefore, the method ignores the sense of terms such as synonyms and antonyms. The earliest known technique for smoothing the term distributions through the use of latent classes is the Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), and it has been shown to improve the performance of a number of information access including text classification (Xue et al., 2008). It is definitely worth trying with our method to achieve classification accuracy from different period of training and test data as high as that from the same time period of these data.

Finally, we evaluated the effect of the method against the number of training documents. Figures 5, 6 and 7 show the results using the test data collected from 1991 - 2000, 2001, and 2010, respectively. As we can see from Figures 5, 6 and 7, the results obtained by TTS were higher than those obtained by kNN and Salles *et al.* methods regardless of the number of training documents. Moreover, when the training and test data are the same time period, the F-score obtained by TTS using

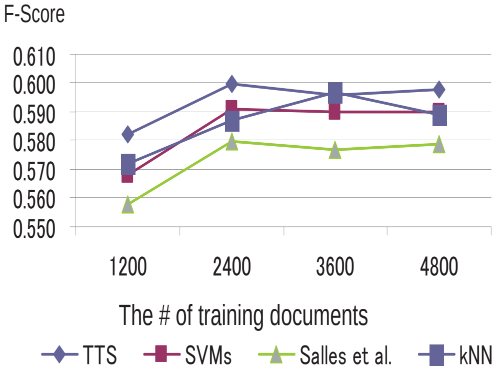

Figure 7: Performance (2010 data)

4,800 training documents and 1,200 documents were 0.821 and 0.804, respectively, and the performance was 1.7% decrease when the training data was reduced. However, those obtained by kNN and Salles *et al.* methods were 2.3% and 2.9% decreases, respectively. The behavior was similar when we used 2001 and 2010 test data. These observations support the effectiveness of our method.

6 Conclusion

We have developed an approach for text classification concerned with the impact that the variation of the strength of term-class relationship over time. We proposed a method of temporal-based term selection for timeline adaptation. The results showed that our method achieved better results than the baselines, kNN and Salles's methods in all of the three types of test data, 1991 - 2000, 2001, and 2010 test data. The result obtained by our method was statistically significant than SVM when the test data (2010) was far from the training data (1991 - 2000), while there was no significant difference between SVM and our method when the period of test data is close to the training data. Moreover, we found that the method is effective for a small number of training documents.

There are a number of interesting directions for future work. We should be able to obtain further advantages in efficacy in our approach by smoothing the term distributions through the use of latent classes in the PLSA (Hofmann, 1999; Xue et al., 2008). We used Japanese newspaper documents in the experiments. For quantitative evaluation, we need to apply our method to other data such as ACM-DL and a large, heterogeneous collection of web content. Temporal weighting function we used needs tagged corpora with long periods of time. The quantity of the training documents affects its performance. However, documents are annotated by hand, and manual annotation of documents is extremely expensive and time-consuming. In the future, we will try to extend the framework by using unsupervised methods *e.g.* Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Wang and McCallum, 2006).

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their valuable comments on the earlier version of this paper. This work was supposed by the Grant-in-aid for the Japan Society for the Promotion of Science (No. 25330255, 26330247).

References

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

W. W. Cohen and Y. Singer. 1999. Context-sensitive Learning Methods for Text Categorization. *ACM Transactions of Information Systems*, 17(2):141–173.

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

G. Folino, C. Pizzuti, and G. Spezzano. 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*, pages 183–188.

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

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

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

D. He and D. S. Parker. 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*, pages 443–452.

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

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

M. G. Kelly, D. J. Hand, and N. M. Adams. 1999. The Impact of Changing Populations on Classifier Performance. In *Proc. of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 367–371.

Y. S. Kim, S. S. Park, E. Deards, and B. H. Kang. 2004. Adaptive Web Document Classification with MCRDR. In *Proc. of 2004 International Conference on Information Technology: Coding and Computing*, pages 476–480.

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

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

R. L. Liu and Y. L. Lu. 2002. Incremental Context Mining for Adaptive Document Classification. In *Proc. of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 599–604.

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

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

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

L. Rocha, F. Mourao, A. Pereira, M. A. Goncalves, and W. M. Jr. 2008. Exploiting Temporal Contexts in Text Classification. In *Proc. of the 17th ACM Conference on Information and Knowledge Management*, pages 26–30.

G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand. 2012. Exponentially Weighted Moving Average Charts for Detecting Concept Drift. *Pattern Recognition Letters*, 33(2):191–198.

T. Salles, L. Rocha, and G. L. Pappa. 2010. Temporally-aware Algorithms for Document Classification. In *Proc. of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314.

M. Scholz and R. Klinkenberg. 2007. Boosting Classifiers for Drifting Concepts. *Intelligent Data Analysis*, 11(1):3–28.

X. Wang and A. McCallum. 2006. Topic over Time: A Non-Markov Continuous-Time Model of Topic Trends. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433.

G. R. Xue, W. Dai, Q. Yang, and Y. Yu. 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*, pages 627–634.

Y. Yang and X. Lin. 1999. A Re-examination of Text Categorization Methods. In *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49.

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