An Empirical Study on Selective Sampling in Active Learning for Splog Detection

Taichi Katayama Takehito Utsuro Yuuki Sato
University of Tsukuba
Tsukuba, 305-8573, JAPAN

Takayuki Yoshinaka
Tokyo Denki University
Tokyo, 101-8457, JAPAN

Yasuhide Kawada
Navix Co., Ltd.
Tokyo, 141-0031, JAPAN

Tomohiro Fukuhara
University of Tokyo, Kashiwa
277-8568, JAPAN

ABSTRACT

This paper studies how to reduce the amount of human supervision for identifying splogs / authentic blogs in the context of continuously updating splog data sets year by year. Following the previous works on active learning, against the task of splog / authentic blog detection, this paper empirically examines several strategies for selective sampling in active learning by Support Vector Machines (SVMs). As a confidence measure of SVMs learning, we employ the distance from the separating hyperplane to each test instance, which have been well studied in active learning for text classification. Unlike those results of applying active learning to text classification tasks, in the task of splog / authentic blog detection of this paper, it is not the case that adding least confident samples peforms best.

Categories and Subject Descriptors

H.3.0 [**INFORMATION STORAGE AND RETRIEVAL**]: General

General Terms

Reliability

Keywords

spam blog detection, SVM, active learning, selective sampling

1. INTRODUCTION

Weblogs or blogs are considered to be one of personal journals, market or product commentaries. While traditional search engines continue to discover and index blogs, the blogosphere has produced custom blog search and analysis engines, systems that employ specialized information retrieval techniques. With respect to blog analysis services on the Internet, there are several commercial and non-commercial services such as *Technorati*[1], *BlogPulse*[2] [3], *kizasi.jp*[3], and *blogWatcher*[4] [13]. With respect to multilingual blog services, *Globe of Blogs*[5] provides a retrieval function of blog articles across languages. *Best Blogs in Asia Directory*[6] also provides a retrieval function for Asian language blogs. *Blogwise*[7] also analyzes multilingual blog articles.

As with most Internet-enabled applications, the ease of content creation and distribution makes the blogosphere spam prone [4, 1, 7, 11, 6]. Spam blogs or splogs are blogs hosting spam posts, created using machine generated or hijacked content for the sole purpose of hosting advertisements or boosting the ranking of target sites. [7] reported that for English blogs, around 88% of all pinging URLs (i.e., blog homepages) are splogs, which account for about 75% of all pings [2]. Based on this estimation, as stated in [1, 10], splogs can cause problems including the degradation of information retrieval quality and the significant waste of network and storage resources. Several previous works [7, 11, 6] reported important characteristics of splogs. [11] reported characteristics of ping time series, in-degree/out-degree distributions, and typical words in splogs found in TREC[8] Blog06 data collection. [7, 6] also reported the results of analyzing splogs in the *BlogPulse* data set. In the context of semi-automatically collecting web spam pages/hosts including splogs, [18] discuss how to collect spammer-targeted keywords to be used when collecting a large number of web spam pages/hosts efficiently. [14] also analyzes (Japanese) splogs based on various characteristics of keywords contained in them.

Along with those analysis on splogs reported in previous works, several splog detection techniques (e.g., [12, 5, 10]) have been proposed. [5] studied features for splog detection such as words, URLs, anchor texts, links, and HTML meta tags in supervised learning by SVMs. As features of SVMs, [10] studied temporal self similarities of splogs such as posting times, post contents, and affiliated links. [12] also

AIRWeb '09, April 21, 2009 Madrid, Spain.
Copyright 2009 ACM 978-1-60558-438-6 ...$5.00.

[1] `http://technorati.com/`
[2] `http://www.blogpulse.com/`
[3] `http://kizasi.jp/` (in Japanese)
[4] `http://blogwatcher.pi.titech.ac.jp/` (in Japanese)
[5] `http://www.globeofblogs.com/`
[6] `http://www.misohoni.com/bba/`
[7] `http://www.blogwise.com/`
[8] `http://trec.nist.gov/`

Table 1: Statistics of Splogs / Authentic Blogs Data Sets

(a) Statistics of Total Data Sets Available

Data Sets	# of splogs	# of authentic blogs	total
Years 2007-2008	768	3318	4086
Years 2008-2009	1445	2459	3904
total	2213	5777	7990

(b) Statistics of Virtual Data Sets for Evaluation in Section 5.2

Years 2007-2008	40×18=720	40×18=720	1440
Years 2008-2009	40×18=720 (40×10=400 used, 360 for training, 40 for evaluation)	40×18=720 (40×10=400 used, 360 for training, 40 for evaluation)	1440 (800 used, 720 for training, 80 for evaluation)

(c) Statistics of Data Sets for Active Learning in Section 5.3

Years 2008-2009	4 for initial training, 1296 for pool, 145 for evaluation	6 for initial training, 2208 for pool, 245 for evaluation	10 for initial training, 3504 for pool, 390 for evaluation

studied detecting link spam in splogs by comparing the language models among the blog post, the comment, and pages linked by the comments.

However, splogs may change year by year. This is partially because text content of splogs is mostly excerpted from other sources such as news articles, blog articles (posts), advertisement pages, and other web texts. Sources of splog contents such as those above may change day by day, and thus, splog contents excerpted from those sources also may change. Furthermore, certain percentage of splogs may be created automatically, where their html structures are automatically generated and their text contents are excerpted from other sources. Such automatic procedures may also change year by year, and hence, it is quite reasonable to suppose that certain characteristics of generated splogs may change year by year. The evaluation results in section 5.2 indirectly support the claim and can be summarized below: the performance of applying Support Vector Machines (SVMs) [17] model to splog detection trained with a Japanese splog data set developed in the years of 2007-2008 [14] against another Japanese splog data set developed in the years of 2008-2009 is relatively damaged compared with that of applying another model trained with a held out data set from the years 2008-2009. Therefore, in order to catch up with such splog changes, it is quite necessary to design a framework for continuously updating splog data sets year by year[9].

In such a framework, one of the most important issues is how to reduce human supervision in continuously updating splog data sets. In machine learning communities and statistical natural language processing communities, minimally supervised approaches such as *active learning* [9, 16, 15] have been well studied. In active learning, certain confidence measures in machine learning frameworks are introduced so as to separate highly confident samples and less confident samples. In previous works on applying active learning frameworks to tasks such as text classification [9, 16, 15], the least confident samples are collected, manually annotated, and added to the training data set. Following those studies in the previous works on active learning, against the task of splog / authentic blog detection, this paper empirically examines several strategies for selective sampling in active learning by SVMs. As a confidence measure of SVMs learning, we employ the distance from the separating hyperplane to each test instance [16], which have been well studied in active learning for text classification.

Unlike those results of applying active learning to text classification tasks, in the task of splog / authentic blog detection of this paper, it is not the case that adding least confident samples peforms best. In total, the strategy of adding least confident samples performs worse than that of adding balanced samples, and even worse than that of randomly selecting samples to be added to the training data set. From this result, we conclude that, when reducing human supervision in continuously updating splog data sets year by year, it is the most important to apply the confidence measure so that novel blog homepages be balancedly sampled in terms of the distance from the separating hyperplane.

2. SPLOGS / AUTHENTIC BLOGS DATA SETS

In this paper, we examine splogs changes over time with two data sets of Japanese splog / authentic blog homepages, where one set is developed in the years of 2007-2008 (from September 2007 to February 2008) [14], and the other set is developed in the years of 2008-2009 (from December 2008 to January 2009). As shown in Table 1, in both data sets, blog homepages are collected according to certain procedures, and then, splog / authentic blog judgement is manually annotated. Here, roughly speaking, splog / authentic blog judgement is based on the criterion below, while the detailed discussion on this criterion is in [14]:

1. If one of the followings holds for the given homepage, then it is mostly[10] *splog.*

   (a) The feature "originally written text" does not hold.

   (b) The feature "originally written text" holds and at least one of the features "links to affiliated sites", "advertisement articles (posts)", or "articles (posts) with adult content" holds.

2. Otherwise, the given homepage is an *authentic blog.*

[9] The notion of *concept drift* [8] and previously studied approaches to detecting it may be closely related to this result.

[10] By "mostly", we mean that it is usually necessary to judge by considering the contents of each blog.

The reason why splog / authentic blog distributions differ between the two data sets is that candidate blog homepages to which splog / authentic blog distinction is annotated are collected according to different procedures. Since it was our first experience of developing splog data set, for the one developed in the years of 2007-2008, we examine characteristics of splogs in advance, and we collect candidate blog homepages which satisfy the requirements below:

i) We collect various keywords including those which are supposed to be frequently included in splogs. And then, for each keyword, we simply collect blog homepages which contain the keyword.

ii) In our observation, the rate of splogs among the blog homepages that contain a given keyword may be higher on the burst date than on other dates. Based on this tendency, we collect blog homepages containing a given keyword on the date with its most frequent occurrence.

iii) Out of the collected blog homepages for a given keyword and on the date above ii), we prefer blog homepages with more posts per day than those with fewer posts per day.

When developing the other data set in the years 2008-2009, on the other hand, we collect candidate blog homepages more efficiently. According to the analysis in [14] as well as our further observation, splog homepages tend to share out-links to affiliated sites. Based on this tendency, in the years 2008-2009, we collect splog / authentic blog homepages which share out-links with splog homepages included in the data set developed in the years 2007-2008[11]. More specifically, first, URLs of the out-links are extracted from the html files of the splog / authentic blog homepages in the data set developed in the years 2007-2008. Then, the set of blacklist URLs is constructed by collecting URLs which satisfy both of the following two requirements:

i) The URL is not included in the html files of any of the training instances of authentic blog homepages.

ii) The URL is included in the html files of the training instances of splog homepages, and its total frequency in the whole training splog homepages is more than one.

About 5,000 blacklist URLs are collected, and finally, for each blacklist URL, candidate blog homepages which include out-link to it are collected. The data set shown in Table 1 is developed from a part of all the collected candidate blog homepages.

3. FEATURES FOR SPLOG DETECTION

This section describes features of SVMs for splog detection. Features described next in this section are evaluated through splog detection performance, and in the evaluation of section 5, only about half of them are manually selected, since the set of those selected features perform best compared with other combinations out of all the features.

3.1 Blacklist/Whitelist URLs

Given the training instances of splog / authentic blog homepages, URLs of the out-links are extracted from their html files. Then, the set of whitelist URLs is constructed by collecting URLs which satisfy both of the following two requirements:

i) The URL is not included in the html files of any of the training instances of splog homepages.

ii) The URL is included in the html files of the training instances of authentic blog homepages, and its total frequency in the whole training authentic blog homepages is more than one.

From the Japanese splog data set developed in the years of 2007-2008 [14] (section 2), about 13,000 whitelist URLs are collected. Next, given a whitelist URL u, the following weight is calculated and is used as a value of the whitelist URLs feature:

$$\log \sum_{u} \left(\begin{array}{c} \text{total frequency} \\ \text{of } u \text{ in the whole} \\ \text{training instances} \\ \text{of authentic blog} \\ \text{homepages} \end{array} \right) \times \left(\begin{array}{c} \text{total} \\ \text{frequency} \\ \text{of } u \text{ in} \\ \text{the test} \\ \text{instance} \end{array} \right)$$

The set of blacklist URLs is also constructed according to a similar procedure (as described in section 2). About 5,000 blacklist URLs are collected, and are evaluated as a feature for splog detection. However, the blacklist URLs feature does not contribute to improving the performance of the best combination of features.

3.2 Noun Phrases

As previously reported in [18, 14], splogs and authentic blogs tend to have word distributions different from each other, and certain types of words may appear in splogs more often than in authentic blogs. In this paper, we introduce a feature for observing occurrences of noun phrases, so that such a difference can be detected. First, body texts of splog / authentic blog posts are extracted, and are morphologically analyzed[12], from which noun phrases are extracted.

Then, given a noun phrase w, based on the following contingency table of co-occurrence frequencies of the whole training instances of splog / authentic blog homepages and w, we estimate correlation of splog / authentic blog homepages and w according to the ϕ^2 statistic between splog homepages and w.

[11] One may argue that candidate blog homepages sharing out-links with splogs taken from the data set of the years 2007-2008 could be biased toward those similar to the ones included the data set of the years 2007-2008. However, as we show in section 5.2, the SVMs model trained with the data set of years 2007-2008 perform worse against splogs / authentic blogs of the years 2008-2009, than against those of the years 2007-2008. Thus, at present, we conclude that our data set of the years 2008-2009 is worthy of further analysis on changes in splogs over time and being examined through splog detection research activities.

[12] The Japanese morphological analyzer ChaSen (`http://chasen-legacy.sourceforge.jp/`) and the lexicon ipadic are used.

Table 2: Selective Sampling Strategies in Active Learning

		Samples to be Selected are with High? or Low? Confidence				
		high	low	high or low	balanced	random
From the Seperating Hyperplance, Samples to be Selected are on Splog side? or Authentic Blog side?	splog	—	—	—	—	Random
	authentic blog	High-Au	—	High/Low-Au	—	
	splog or authentic blog	—	Low-Sp/Au	—	Balanced-Sp/Au	

	w	$\neg w$
the whole training instances of splog homepages	$freq(\text{splog}, w) = a$	$freq(\text{splog}, \neg w) = b$
the whole training instances of authentic blog homepages	$freq(\text{authentic blog}, w) = c$	$freq(\text{authentic blog}, \neg w) = d$

$$\phi^2(\text{splog}, w) = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

From the Japanese splog / authentic blog data sets developed in the years of 2007-2008, about 462,894 noun phrases are collected. Next, the following weight is calculated from the whole training instances of splog / authentic blog homepages and is used as a value of the splog noun phrase feature:

$$\log \sum_{w} \phi^2(\text{splog}, w) \times \left(\begin{array}{c} \text{total frequency of } w \\ \text{in the test instance} \end{array} \right)$$

3.3 Noun Phrases in Anchor Texts and Linked URLs

Another type of features which are more specific than the blacklist/whitelist features and noun phrase features, but are more effective in splog detection is that of (loose) tuple of noun phrases in anchor texts and their linked URLs. In order to introduce this feature, given a noun phrase w and a splog / authentic blog homepage s, first we define the *frequencies* $AncfB(w, s)$ and $AncfW(w, s)$ of a noun phrase w in s:

$$AncfB(w, s) = \left(\begin{array}{c} \text{\# of times of } w \text{ in } s \text{ s.t.} \\ w \text{ is included in an anchor text} \\ \text{of an out-link to a blacklist URL or} \\ \text{a post of splog homepage included} \\ \text{in the training data set.} \end{array} \right)$$

$$AncfW(w, s) = \left(\begin{array}{c} \text{\# of times of } w \text{ in } s \text{ s.t.} \\ w \text{ is included in an anchor text} \\ \text{of an out-link to a whitelist URL or} \\ \text{a post of authentic blog homepage} \\ \text{included in the training data set.} \end{array} \right)$$

Then, the set of *splog anchor text noun phrases out-linked to blacklist URLs* whose total frequency throughout the whole training splog homepages $\sum_{s} AncfB(w, s)$ is more than one is constructed, where from the Japanese splog data set developed in the years of 2007-2008, about 2,000 anchor text noun phrases are collected. Next, given a *splog anchor text noun phrase out-linked to blacklist URLs* w, the following weight is calculated and is used as the value of a feature named *anchor text noun phrase out-linked to blacklist URLs* for a test instance blog homepage t:

$$\log \sum_{w} \Big(\sum_{\substack{\text{training} \\ \text{splog} \\ \text{homepage} \\ s}} AncfB(w, s) \Big) \times AncfB(w, t)$$

In a similar procedure, the set of *splog anchor text noun phrases out-linked to whitelist URLs* whose total frequency throughout the whole training splog homepages $\sum_{s} AncfW(w, s)$ is more than one is constructed, where from the Japanese splog data set developed in the years of 2007-2008, about 320 anchor text noun phrases are collected. Next, given a *splog anchor text noun phrase out-linked to whitelist URLs* w, the following weight is calculated and is used as the value of a feature named *anchor text noun phrase out-linked to whitelist URLs* for a test instance blog homepage t:

$$\log \sum_{w} \Big(\sum_{\substack{\text{training} \\ \text{splog} \\ \text{homepage} \\ s}} AncfW(w, s) \Big) \times AncfW(w, t)$$

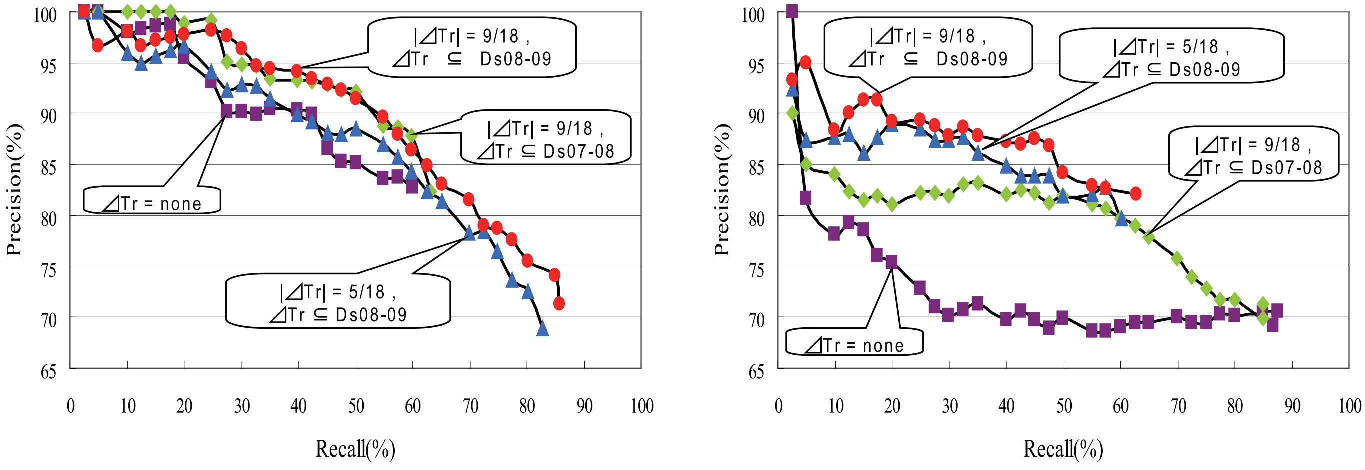
3.4 Link Structure

In addition to the features introduced so far, we also examine features for representing link structures such as the out-degree, the maximum number of out-links from a blog homepage to any one URL, and the number of mutual links to any other blog homepages. However, any of those features contribute to improving the performance of the best combination of features.

4. ACTIVE LEARNING FOR SPLOG DETECTION

4.1 Splog Detection by SVMs

As a tool for learning SVMs, we use TinySVM (`http://chasen.org/~taku/software/TinySVM/`). As the kernel function, we compare the linear and the polynomial (2nd



(a) Splog Detection

(b) Authentic Blog Detection

Figure 1: Evaluation Results of Splog / Authentic Blog Detection (Training: $\frac{9}{18}$ of Ds07-08 (Years 2007-2008) + ΔTr, Evaluation: $\frac{1}{18}$ of Ds08-09 (Years 2008-2009))

order) kernels, where the 2nd order kernels perform better and we show results with the 2nd order kernels in this paper.

4.2 A Confidence Measure

As a confidence measure of SVMs learning, we employ the distance from the separating hyperplane to each test instance [16][13]. More specifically, we introduce two lower bounds LBD_s and LBD_{ab} for the distance from the separating hyperplane to each test instance, where LBD_s is for the test instances judged as *splogs*, while LBD_{ab} is for those judged as *authentic blogs*. If a test instance x is judged as a splog, but its distance from the separating hyperplane is not greater than LBD_s, then the decision is not regarded as confident and is rejected. In the case of x being judged as an authentic blog, the lower bound LBD_{ab} is considered in a similar fashion.

4.3 Selective Sampling Strategies in Active Learning

This paper empirically examines several strategies for selective sampling in active learning by SVMs. In the experimental evaluation, we start from randomly selected 10 initial training instances (4 splog homepages and 6 authentic blog homepages) and 390 instances for evaluation. As shown in Table 1 (c), with 3504 unlabeled instances as a pool, at each step of active learning, 4 unlabeled instances are selected out of the pool according to a certain strategy, manually annotated whether splog or authentic blog, and then added to the training data set[14]. Active learning cycles continue until 1,000 instances are added to the training data set[15].

At each step of active learning, the current model for splog / authentic blog detection judges with certain confidence whether each of unlabeled instances in the pool is splog / authentic blog. The results of those judgements are represented as the location of each unlabeled instance in the feature vector space, where the space is divided into the splog side and the authentic blog side by the separating hyperplane. The confidence of the judgments by the current model is represented the distance from the separating hyperplane. The high the confidence is, the larger the distance is from the separating hyperplane.

In our experimental evaluation of various strategies for selective sampling, the whole strategies are shown in their acronyms in Table 2. Here, we consider the following two factors:

- The confidence that is assigned by the current model to each unlabeled instance. In previous works on active learning, the least confident instances are collected to be added to the training data set. In this paper, we examine the following four strategies: i) the most confident instances are selected ("high" in Table 2), ii) the least confident ones are selected ("low" in Table 2), iii) both the most confident ones as well as the least confident ones are selected ("high or low" in Table 2), iv) instances to be selected are balanced in terms of confidence, i.e., the distance from the separating hyperplane ("balanced" in Table 2).

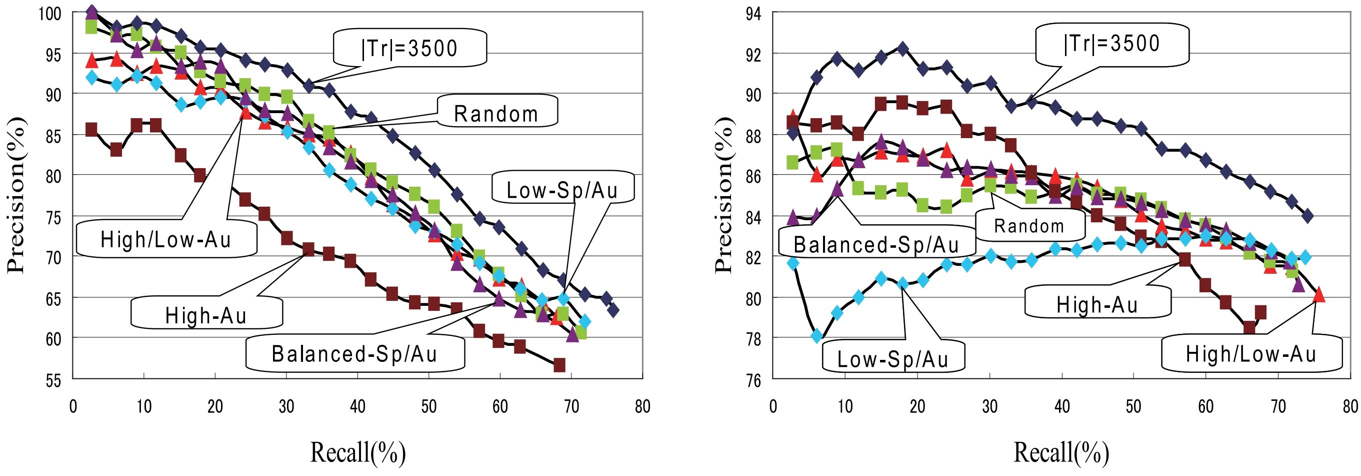
- Within the feature vector space divided by the separating hyperplane, unlabeled instances are selected from the splog side, or from the authentic blog side, or from both sides.

As shown in Table 2, any pair of those two factors is considered and experimentally evaluated. Furthermore, we also compare the strategy of randomly selecting unlabeled instances in the pool with the strategies we designed above.

[13] [16] studied this measure in the context of *active learning* [9, 16, 15], one of major minimally supervised approaches in machine learning communities and statistical natural language processing communities, where, in active learning framework, the least confident samples are collected, manually annotated, and added to the training data set.

[14] We ignore training instances for which values of all the features are zero.

[15] At present, we have not examined the issue of stopping criterion.



(1) Splog Detection

(2) Authentic Blog Detection

Figure 2: Evaluation Results of Active Learning: Models after Adding 1,000 Training Samples

For each of the strategies "high", "low", "high or low", and "balanced", we compare the performance of the three alternatives in the orthogonal factor, namely, from the splog side, from the authentic blog side, and from both sides. In Table 2, we only list the acronyms of the best performing strategy for each of "high", "low", and "high or low".

5. EXPERIMENTAL ANALYSIS

5.1 Evaluation Measures

Throughout this paper, the performance of splogs / authentic blogs detection are shown with plots of recall and precision. Here, recall / precision of detecting splogs and detecting authentic blogs are measured separately. The performance curve of splog detection is plotted by varying the lower bound LBD_s in splog side of the data space, while that of authentic blog detection is plotted by varying the lower bound LBD_{ab} on authentic blog side of the data space.

5.2 Estimating Changes in Splogs over Time through Performance of Splog Detection

In the first experimental evaluation, we compare the two data sets developed in the years 2007-2008 and 2008-2009 through performance of splog / authentic blog detection. Here, we trained two classifiers, one with the data set developed in the years 2007-2008, while the other with the mixture of the two data sets[16]. Then, we evaluate the two classifiers against a held out data taken from the data set developed in the years 2008-2009.

In Figure 1, the plots labeled with "ΔTr=none" and "$|\Delta Tr| = \frac{9}{18}$, $\Delta Tr \subseteq$ Ds07-08" indicate the performance of those trained with the data set from the years 2007-2008, where the training data size for the former is half of that for the latter. Those with "$|\Delta Tr| = \frac{9}{18}$, $\Delta Tr \subseteq$ Ds08-09" and "$|\Delta Tr| = \frac{5}{18}$, $\Delta Tr \subseteq$ Ds08-09" indicate the performance of those trained with the mixture of the two data sets. Considering the performance of both splog and authentic blog detection, the best performance are obtained when the classifier is trained with the mixture of the two data sets. This is obviously because both (a part of) training and evaluation data are from the data sets of the same period. Especially, the classifier for authentic blog detection tend to judge splogs of the years 2008-2009 as authentic blogs. Furthermore, with more training data from the years 2008-2009, the classifier performs better. This result clearly indicates that (at least) splogs change over time (in this case, about one year), and their changes are somehow not very easy for the SVM classifier trained in the past to catch up with.

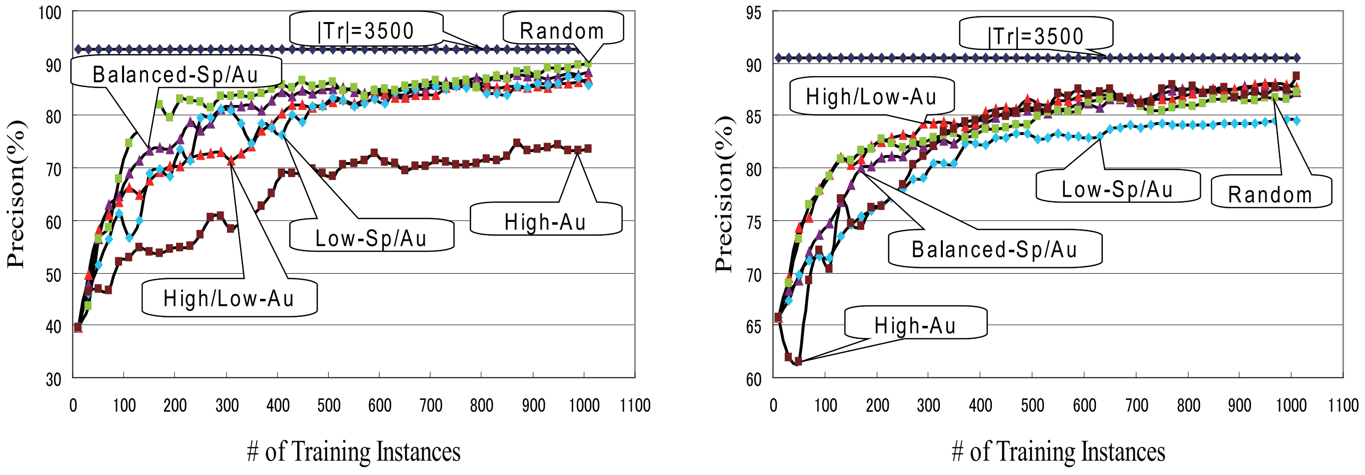
5.3 Selective Sampling in Active Learning for Splog Detection

For each strategy in Table 2, Figure 2 plot the recall / precision curves after adding 1,000 instances to the training data set. Figure 3 also plots the learning curve of the maximum precisions at the point of the recall as 30%. In those figures, the plots with the label "$|Tr| = 3500$" indicate that the whole 3504 instances in the pool.

Among those strategies in the figures, "Balanced-Sp/Au", "Random", and "High/Low-Au" perform comparably well. The strategies of adding the most/least confident instances perform worse. From this result, we conclude that, when reducing human supervision in continuously updating splog data sets year by year, it is the most important to apply the confidence measure so that novel blog homepages be balancedly sampled in terms of the distance from the separating hyperplane. Also from Figure 3, both splog and authentic blog detection performance do not seem saturated after adding 1,000 training instances.

Finally, Figure 4 plots changes in the numbers of support vectors. In SVMs learning, it is known that only the support vectors have effect on deciding the position of the separating

[16] The specifications of the training and evaluation data sets are given in Table 1 (b). Out of the total splogs / authentic blogs data sets available in Table 1 (a), the training data sets of the same size are constructed both from the years 2007-2008 and from the years 2008-2009. Here, note that within both the training and evaluation data sets, the distribution of splogs and authentic blogs is evenly 50% and 50%.



(1) Splog Detection

(2) Authentic Blog Detection

Figure 3: Evaluation Results of Active Learning: Changes in Maximum Precisions with Recall as 30%

hyperplane, and the number of support vectors can be regarded as the complexity of the learning task. As can be seen from the result of Figure 4 (a), the number of support vectors continue to increase. The strategies such as “Low-Sp/Au” and “High-Au” with relatively low performance are with less support vectors, and thus seem to fail in collecting certain numbers of effective training instances which are successfully collected by other strategies. From the results in Figure 4 (b) and (c), the rate $| SVm \bigcap SVall | / | SVm |$ starts to be saturated, while the rate $| SVm \bigcap SVall | / | SVall |$ continues to increase. This result is contrastive compared with the one in Figure 2, where the performance after adding 1,000 training instances is relatively close to the plots of “$| Tr |= 3500$”. This indicates that although the recall / precision of splog / authentic blog detection continue to improve relatively slowly, the number of support vectors increase much more rapidly.

6. CONCLUSION

This paper studied how to reduce the amount of human supervision for identifying splogs / authentic blogs in the context of continuously updating splog data sets year by year. Following the previous works on active learning, against the task of splog / authentic blog detection, this paper empirically examined several strategies for selective sampling in active learning by Support Vector Machines (SVMs). Unlike those results of applying active learning to text classification tasks, in the task of splog / authentic blog detection of this paper, it is the most effective to add samples that are balanced in terms of the distance from the separating hyperplane. Future works include, within the framework of active learning, introducing other features such as ping time series that are studied in the previous works [11, 7, 6]. We are also working on updating splog / authentic blog data set through the manual procedure by annotators, where instances to be shown to the annotators are automatically selected considering their distance from the separating hyperplane. The result of this work will be reported in the near future.

7. REFERENCES

[1] *Wikipedia, Spam blog.* `http://en.wikipedia.org/wiki/Spam_blog`.

[2] *Wikipedia, Ping (blogging).* `http://en.wikipedia.org/wiki/Ping_(blogging)`.

[3] N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated trend discovery for Weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[4] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc. 1st AIRWeb*, pages 39–47, 2005.

[5] P. Kolari, T. Finin, and A. Joshi. SVMs for the Blogosphere: Blog identification and Splog detection. In *Proc. 2006 AAAI Spring Symp. Computational Approaches to Analyzing Weblogs*, pages 92–99, 2006.

[6] P. Kolari, T. Finin, and A. Joshi. Spam in blogs and social media. In *Tutorial at ICWSM*, 2007.

[7] P. Kolari, A. Joshi, and T. Finin. Characterizing the splogosphere. In *Proc. 3rd Ann. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[8] L. I. Kuncheva. Classifier ensembles for detecting concept change in streaming data: Overview and perspectives. In *Proc. 2nd Workshop SUEMA 2008 (ECAI 2008)*, pages 5–10, 2008.

[9] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proc. 17th SIGIR*, pages 3–12, 1994.

[10] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *Proc. 3rd AIRWeb*, pages 1–8, 2007.

[11] C. Macdonald and I. Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow, Department of Computing Science, 2006.

[12] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proc. 1st AIRWeb*, 2005.

[13] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining Japanese weblogs. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 320–321. ACM Press, 2004.

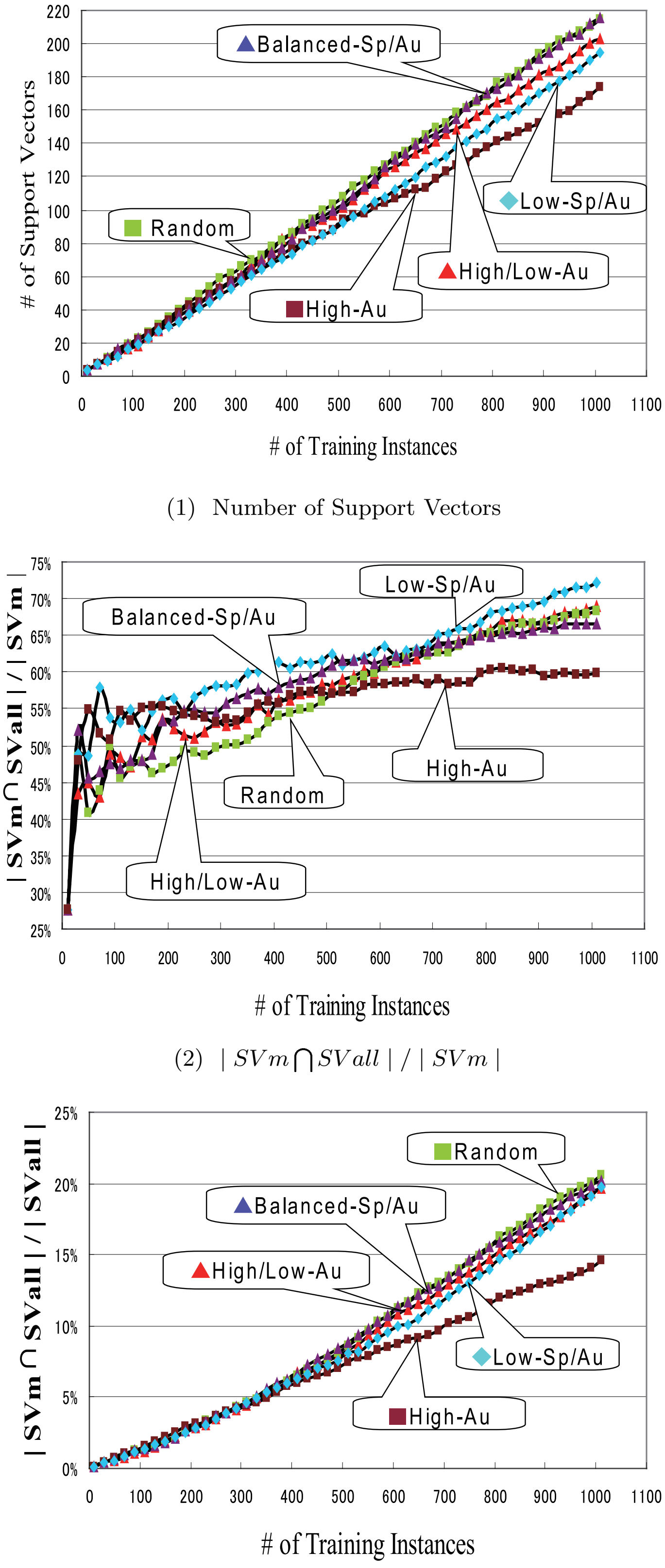
[14] Y. Sato, T. Utsuro, T. Fukuhara, Y. Kawada, Y. Murakami, H. Nakagawa, and N. Kando. Analysing features of Japanese splogs and characteristics of keywords. In *Proc. 4th AIRWeb*, 2008.

[15] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *Proc. 17th ICML*, pages 839–846, 2000.

[16] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proc. 17th ICML*, pages 999–1006, 2000.

[17] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[18] Y. Wang, M. Ma, Y. Niu, and H. Chen. Spam double-funnel: Connecting web spammers with advertisers,. In *Proc. 16th WWW Conf.*, pages 291–300, 2007.



(1) Number of Support Vectors

(2) $|\, SVm \bigcap SVall \,| \,/\, |\, SVm \,|$

(3) $|\, SVm \bigcap SVall \,| \,/\, |\, SVall \,|$

Figure 4: Changes in Number of Support Vectors (SVm: Set of Support Vectors at m-th Active Learning Cycle, $SVall$: Set of Support Vectors after adding the Whole 3504 Instances in the Pool)