Categorizing Blogger's Interests Based on Short Snippets of Blog Posts

Jiahui Liu, Larry Birnbaum, Bryan Pardo
Northwestern University
2133 Sheridan Road, Evanston, IL, 60201, USA
j-liu2@northwestern.edu, {birnbaum, pardo}@cs.northwestern.edu

ABSTRACT
Blogs have become an important medium for people to express opinions and share information on the web. Predicting the interests of bloggers can be beneficial for information retrieval and knowledge discovery in the blogosphere. In this paper, we propose a two-layer classification model to categorize the interests of bloggers based on a set of short snippets collected from their blog posts. Experiments were conducted on a list of bloggers collected from blog directories, with their snippets collected from Google Blog Search. The results show that the proposed method is robust to errors in the lower level and achieve satisfactory performance in categorizing blogger's interests.

Categories and Subject Descriptors
I.5.2 [**Pattern Reorganization**]: Design Methodology-*Classifier design and evaluation*.

General Terms Algorithms, Experimentation, Performance.

Keywords Blogger, interests, categorization

1. INTRODUCTION
As an important form of online publishing for common internet users, blogs have emerged as a dynamic and diversified medium for information creation, distribution and accumulation. Bloggers who are interested in certain domains maintain blog sites to publish news, opinions and ideas about the domains of their interest. Identifying and categorizing the interests of bloggers can be valuable for information retrieval and knowledge discovery from blogs.

The blog posts published by bloggers provide important clues for predicting their interests. However, direct categorization of all the texts written by a blogger cannot produce accurate prediction. This is mainly due to two reasons. First, blog articles are written in an informal erratic style. Bloggers sometimes even invent new words and grammars to express themselves idiosyncratically. Second, bloggers do not confine themselves to one topic [2]. Therefore, the mixture of all the posts by a blogger is a multi-topic and noisy text document that is difficult to classify.

To address these challenges, we propose a two-layer classification model to categorize blogger's interests. In the first layer, a text classifier is trained to predict the probability of a blog post belonging to a domain category. Although the classification of individual post is not perfect, the categorizations of multiple posts of a blogger provide important information to predict the overall interests of that blogger. In the second layer, we derive features from the set of categorization probabilities of the posts written by a blogger. Those features are used to categorize the interests of the blogger. By incorporating the membership information of blog posts regarding all the categories, the second layer classifiers are able to learn the topical correlations among these categories.

We experiment with the proposed model using a collection of bloggers compiled from blog directories, with blog post snippet retrieved from Google Blog Search. As we expected, classification of short snippets in the first layer is not very accurate. But the features derived from the set of probabilities for all the snippets are meaningful and useful for predicting the overall interests of bloggers. Categorization of bloggers' interests achieves F1 measure of 0.845 by microaveraging over all the categories.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10

2. THE PROPOSED TECHNIQUE
In this paper, we use short snippets of a blogger's posts to characterize their interests. Using the snippets eliminates the need to download the full web page. Snippets are also faster to process than full page, enabling real time processing, which is especially critical for web applications. For each blogger b, we collect a set of most recent blog posts written by b, denoted as $P(b) = \{p_1, p_2, ..., p_n\}$. For each post p_i, we extract a short snippet $s_i = Snippet(p_i)$. A snippet consists of the title and the first few sentences of a blog post, containing about 40 words. $S_b = \{s_1, s_2, ..., s_n\}$ is the set of snippets collected for blogger b. The goal is to categorize the interests of b into one or multiple classes, drawn from a set of classes, $C = \{c_1, c_2, ..., c_m\}$.

The proposed technique addresses this task with a two-layer classification model. In the first layer, the classifiers produce a probability estimate $p(c_j \mid s_i)$ for each post snippet s_i, which is the probability that snippet s_i belongs to category c_j. In the second layer, we derive features from the categorization probabilities for all the snippets written by blogger b and use these features to predict the interests of b.

2.1 Categorizing Snippets of Blog Posts
To build text classifiers of snippets, we take the stemmed content words of snippets as features, with stop words removed. For each category c_j, we selected the most predictive 2000 stemmed words according to Information Gain [6].

To categorize the snippets, we use the sequential minimal model (SMO) [3], which has been shown to be efficient and effective for text classification. The output of SVM is fit to a sigmoid model to derive a proper probability estimate of membership [4].

We create a SVM for each category c_j . Snippet s_i is processed by each SVM with a sigmoid model, resulting in an *m*-value vector, where the *j*th feature is the probability that snippet s_i belongs in category c_j.(note that the categories are not mutually exclusive.)

$$s_i = \langle p(c_1 \mid s_i), \ldots p(c_j \mid s_i), \ldots p(c_m \mid s_i) \rangle \quad (1)$$

2.2 Categorizing Blogger's Interests

Categorizations of a blogger's snippets (equation 1) provide important clues about a blogger's interests. Features are derived from the categorization of snippets to train classifiers of bloggers. For each category c_j, we take all the probability estimates $p(c_j \mid s_i)$ for $s_i \in S_b$. The set of probability estimates is

$$E_j = \{ p(c_j \mid s_i), \ldots p(c_j \mid s_i), \ldots p(c_j \mid s_n) \} \quad (2)$$

E_j shows how much a blogger writes about category c_j. The probability estimates in the E_j is binned and placed into a histogram. For each category c_j, we divide the [0, 1] range into K intervals and compute the *K*-element distribution of snippets in S_b according to $p(c_j \mid s_i)$. We denote the *k*th element in the distribution for the *j*th category d_k^j. d_k^j is the proportion of snippets in S_b with $p(c_j \mid s_i)$ falling in the *k*th interval. Formally, d_k^j is computed by Equation 3.

$$d_k^j = \frac{|S_k|}{|S_b|}, \text{ where } S_k = \left\{ s \mid s \in S_b, p(c_j \mid s) \in \left[\frac{k-1}{K}, \frac{k}{K} \right) \right\} \quad (3)$$

We also calculate the mean and variance of $p(c_j \mid s_i)$ for each category c_j. These are denoted d_{mean}^j and d_{var}^j. The proportions, mean and variance form a group of features D_j for category c_j.

$$D_j = \{ d_1^j, \ldots, d_K^j, d_{mean}^j, d_{\text{var}}^j \} \quad (4)$$

D_j characterize how much blogger *b* writes about category c_j. However, knowing that a blogger wrote some articles about category c_j is not enough to predict that she is interested in c_j. This is because of the multi-topic nature of blogs [2]. To characterize a blogger's interests, we use all the features derived for all of the categories. A blogger *b* is encoded as the union of D_j for $C = \{c_1, c_2, \ldots, c_m\}$, as shown in (5)

$$b = \{ D_1, \ldots, D_j, \ldots D_m \} \quad (5)$$

To categorize bloggers' interests, we train the second layer of classifiers using the derived features shown in (5). We experimented with a number of machine learning algorithms, including SMO [3], nearest neighbor [1], and neural network with one hidden layer. Our experiment shows that the neural network achieves the highest F1 measure.

3. EXPERIMENTS

For our experiments, we collected 4,428 blog sites from BlogCatalog and the blog section of Yahoo directory. The sites are in 8 major categories: *art*, *business*, *education*, *health*, *law*, *politics*, *religion* and *technology* in the blog directories. In our experiment, we assume that each blog site is owned by a single blogger. We labeled each blogger with the categories assigned to their blog sites in the blog directories. Blog snippets for the bloggers were collected using Google Blog Search. We queried the blog search engine with the URL of each blog site and collected the most recent 30 (or less) results for each blogger. The title and the search result summary returned by the search engine are used together as the snippet. Altogether we collected 86,598 blog post snippets for the 4,428 bloggers. In our experiment on the proposed two-layer classification model, we needed twos separate datasets for classifies in each layer. We randomly divided the bloggers into two halves. The snippets retrieved for the first half of bloggers were used to train the first layer classifiers for blog snippets. Using the snippet classifiers, we evaluate the second layer classifiers for bloggers on the second half of bloggers using 10-fold cross-validation.

We implemented the two-layer classification model describe above using the Weka package [5]. To evaluate the classifiers in each layer, we used the conventional precision, recall and F1 measures. We computed the micro-averaged values for the three measures, which combines the performance of individual categories, weighted by the number of instances in the categories.

Table 1 Performance of two-layer classification

	Precision	**Recall**	**F1 measure**
snippets classification	0.717	0.416	0.526
blogger classification	0.870	0.822	0.845

As shown in table 1, the classification of the first layer classifiers of short snippets is not very accurate. However, in the second layer classification of blogger's interests, the classifiers achieved micro-level F1 of 0.845 with neural networks of one hidden layer with 8 nodes. The experiment shows that the second layer classifier is robust to the errors made in the first layer. In other words, although the first layer's accuracy is low, it is sufficient for making predictions in the second layer.

4. CONCLUSION

In this paper, we propose a two-layer classification model for categorizing blogger's interests based on short snippets of their blog posts. In the first layer, we predict the probability of a snippets belonging to each category. In the second layer, we derive features from the set of probabilities for snippets written by a blogger and use those features to categorize the blogger's interests. Although short and noisy blog post snippets are hard to classify, the two-layer classification model has been shown to be robust to the errors made in the lower level and achieve satisfactory performance in categorizing blogger's interests.

5. REFERENCES

[1] Martin, B. 1995. "Instance-Based learning: Nearest Neighbor With Generalization". Hamilton, New Zealand.

[2] Pew Internet and the American Life Project. 2006 http://www.pewinternet.org/PPF/r/186/report_display.asp.

[3] Platt, J. 1998. "Machines using Sequential Minimal Optimization". In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning,

[4] Platt, J. C. 1999. "Probabilities for SV machines". In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, Advances in Large Margin Classifiers. MIT Press.

[5] Witten, I. H. and Frank, E. 2005 "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.

[6] Yang, Y., Pedersen J.P. 1997. "A Comparative Study on Feature Selection in Text Categorization". In ICML 1997.