Spectrum: Retrieving Different Points of View from the Blogosphere

Jiahui Liu, Larry Birnbaum, and Bryan Pardo

Northwestern University
Intelligent Information Laboratory
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 publish opinions and ideas on the web. Bloggers with interest and expertise in specific domains (e.g., politics, or technology) often create and maintain blogs to publish news, opinions and ideas about those domains. In this paper, we present Spectrum, a novel blog search system that enables users to search for different points of view related to a topic from the blogosphere. Given a topic, Spectrum retrieves blog posts from bloggers with interests and expertise in various domains, enabling users to browse and compare the opinions related to different aspects of the topic. To identify bloggers in a domain category, we propose a two-layer classification model that predicts bloggers' interests based on short snippets of posts by the blogger and posts citing the blogger. The model characterizes the recurrent interests of bloggers and the importance of the bloggers in the domain. Experiments were conducted on a list of bloggers collected from blog directories, with their snippets collected from Google Blog Search. Categorization of bloggers' interests achieves precision of 88.4% and recall of 84.5% by micro-averaging over all the categories, outperforming a baseline algorithm which directly classifies the bloggers' snippets. We further apply this multi-perspective blog search to explore the ecological relationship between news and blogs. The system aggregates recent popular news stories and then automatically aggregates different points of view about those news stories in the blogosphere.

Introduction

Blogs have emerged as an important form of online publishing for internet users, and a rich and diversified resource for personal opinions and ideas. Individuals and organizations alike are interested in information from the blogosphere related to a variety of topics. While blog search shares some features with general web search, it is distinct in terms of user information goals and the personal publishing nature of blogs.

To design information systems that help users find useful and interesting information in the blogosphere, it is critical to understand user needs in blog search. Mishne and de Rijke (2006) conducted extensive query log analysis of a blog search engine. Their analysis shows that blog searches have different intents than general web searches. Specifically, users of blog search engine are mainly interested in *opinions* on current news events and thoughts on general topics, such as "stock trading," "gay rights," and "Islam."

By the self-publishing nature of blogs, ideas and opinions in blogs are biased towards the interests of the bloggers. For example, the controversial issue of abortion has multiple aspects, such as health, law, religion, etc. Bloggers who are concerned with different aspects of the issue will have significantly different points of view about it. Current blog search engines (e.g., Google Blog Search or Technorati) which enable users to find blog posts relevant to a topic present their results as a list of posts. In such a list, the characteristics and concerns of the bloggers are unclear to the user. On the other hand, many blog directories have been created (e.g., BlogCatalog and Bloghub) to help users find bloggers with recurring interests in particular domains. However, users cannot search for posts related to a topic in these blog directories. Furthermore, creating and maintaining the directories require a great amount of manual effort. It is hard to keep the directories up-to-date with the rapid changes in the blogosphere.

In this paper, we present *Spectrum*, a novel blog search system that enables users to find the blog posts written by bloggers with interests and expertise in different domains, such as business, politics, technology, and so on. In addition to retrieving blog posts related to a topic, Spectrum filters and categorize the blog search results according to the domain interests of the bloggers. The system automatically identifies important bloggers in a list of topic domains. User's blog search results are presented according to the domain interests of bloggers, allowing

users to compare the opinions of bloggers with different concerns.

To identify the domain interests of bloggers, Spectrum utilizes two kinds of information: the posts written by the bloggers, and the posts citing those bloggers. The bloggers' own posts reveal their intrinsic interests, while citation posts provide information about the importance of the bloggers in those domains. To enable fast processing, the system uses the snippets of blog posts. We propose a two-layer classification model to categorize bloggers' interests with bloggers' snippets and their citation snippets. Our experiment demonstrated that the two-layer model is robust to the noise in heterogeneous blog writing and effective in identifying the main interests of bloggers.

The user analysis conducted by Mishne and de Rijke (2006) shows that users of blog search engines are particularly interested in discussions about current news events. To explore the ecological relationship between news and blogs, we apply the multi-perspective blog search system in the context of news reading which automatically aggregates different points of view about current news stories in the blogosphere.

Related Work

The popularity of blogs has triggered much research in characterizing them in recent years. Durant and Smith (2006) explored techniques to predict the political orientation (i.e. liberal or conservative) of political blog posts. Ni *et al.* (2007) investigated machine learning methods to classify informative and affective articles in blogs. They proposed that blogs containing more informative articles are of higher quality. A blog search engine is presented in (Ni *et al.*, 2007) that allows users to adjust their search along the dimension of informative versus affective. Our work explores blogs along the dimension of bloggers' interests. Furthermore, our system differs from their work in that it categorize at the level of bloggers instead of individual articles.

Some research has been reported that characterizes various properties of internet users in general or bloggers in particular. Hu *et al.* (2007) proposed an approach to predicting users' age and gender based on browsing behavior. Their approach is similar to ours in that they first predict the age and gender tendency of the web pages browsed by a user and then categorize the user according to the predictions. However, as opposed to demographic predictions, the classes in bloggers' interest categorization are not exclusive. A person can only be either male or female, but a blogger can be an expert in both law and economics. Our second layer classifiers take into account the correlation between different categories and allow a blogger to be categorized into multiple classes. In terms of categorizing the properties of bloggers, Schler *et al.* (2006) utilized stylistic and content-based features to predict bloggers' ages and genders; Oberlander and Nowson (2006) report on the task of classifying bloggers' personalities from their posts. In addition to content analysis, other research explores link structure in the blogosphere to characterize bloggers. For instance, Bhagat *et al.* (2007) proposed a method to infer demographic information about bloggers, including age, gender and location, from a set of labeled bloggers in the linked graph. Efron (2004) utilized co-citation information to estimate political orientations of blog sites as well as other web sites. Our work is distinctive from theirs with respect to both our problem of classifying bloggers' interests and the two-layer classification model we propose to make more accurate aggregate predictions based on imperfect lower-level predictions. Based on the two-layer model, we also present another way for using the cross-linking among blogs to categorize bloggers.

Another thread of research pertaining to our work is author topic modeling that infers the relevant topics of authors from large text corpora using unsupervised learning techniques. Steyvers *et al.* (2004) extended a probabilistic topic model to include authorship information and experimented on the CiteSeer digital library. McCallum *et al.* (2005) proposed the Author-Recipient-Topic (ART) model that captures both the interaction structures in social networks and the language content of the interactions. Author topic models are useful for discovering topics in large corpora, clustering authors sharing similar topics and predicting their roles in social networks. In this paper, we are targeting a somewhat different problem, categorizing bloggers' interests based on documents retrieved in real time.

The blogosphere is constantly and rapidly changing, with bloggers joining and leaving, and new articles being posted all the time. We adopt supervised learning techniques for our task. Classifiers learned from a limited amount of training data are used to make predictions in real time about new bloggers with newly created posts.

Part of the work described here includes, as mentioned earlier, a system to retrieve and aggregate different points of view about current news from blogs. There have been some studies about the ecological relationship between news and blogs. Cointet et al. (2007) studied the topic correlation between blogs and news websites. Ikeda et al. (2006) proposed methods to automatically link news articles to blogs that refer to them. BLEWS, developed by Gamon et al. (2008), utilized blogs to provide contextual information for political news articles to in order to gauge the popularity of and sentiments about news topics. The system proposed in this paper explores the relationship between news and blogs at a finer granularity. In addition to finding blogs related to certain topics, the system categorizes them according to the domains of bloggers' interests, enabling users to browse opinions from different perspectives.

Retrieving Different Points of View

To explore the different points of view available in the blogosphere, we present Spectrum, a multi-perspective blog search system that helps users find and browse

interesting blogs. The system allows users to search for blog posts reflecting different aspects of topics and compare the opinions of bloggers with different concerns.

Figure 1 shows the query interface for Spectrum. In addition to query terms, users can also specify the domain they are interested in. The system retrieves posts related to the user's query, and then filters and categorizes the blog search results according to the characteristics of their corresponding authors. If the author is identified as blogger with recurrent interests in the domain selected by the user, the post written by the author is listed in that domain.

Figures 2 and 3 show the result pages for the query "abortion" in categories of religion and law respectively. The user can click on different categories to view the results in those categories. As shown in Figure 3, posts from religious blog sites discuss "abortion" in the context of various religious beliefs. Blog results in other categories present different perspectives about this controversial issue. For example, law bloggers (shown in Figure 2) discuss legislation related to abortion from a legal point of view. In the domain of health care (not shown), bloggers post practical information about abortion choices. Organizing the blog results in different categories enables

Figure 1 Query interface of Spectrum

users to compare the points of view of people with different interests in the same issues.

In addition to finding blog posts in different categories, the system also helps the user find bloggers who are interested in those particular domains. Unlike blog directories, the list of bloggers is dynamically created in response to the user's query. It contains not only the popular and established sites in the blogosphere, but also the sites that are recently created and less well known. Identifying those bloggers concerned with particular

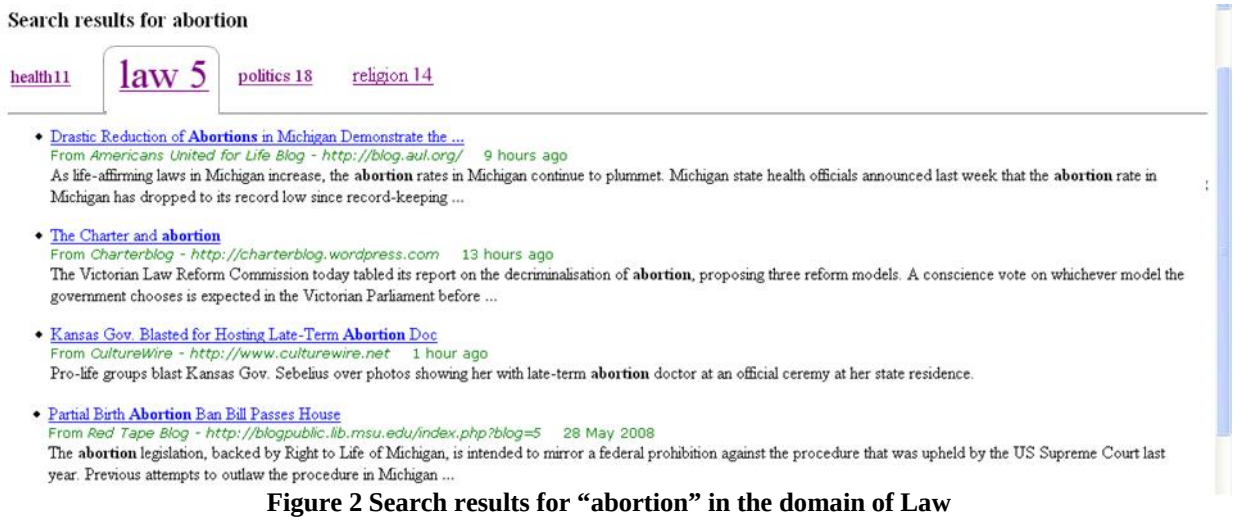

Figure 2 Search results for "abortion" in the domain of Law

Figure 3 Search results for "abortion" in the domain of Religion

domains can be an interesting experience for blog search and facilitate community building in the blogosphere.

Spectrum is implemented as a meta-search system. Figure 4 shows the architecture of the system. The system submits the query string to Google Blog Search and collects the results returned. For each result, the system identifies the URL of the blog site and collects a set of post snippets published on that site and another set of post snippets that cite the blogger. Based on these two set of snippets, the system predicts the main interests of the blogger. If the system does not detect consistent interests in a blogger's posts, that blogger is filtered out. On the other hand, if the blogger's interests do not match the user's interests, the blogger is filtered out as well. Otherwise, the results are organized into the selected categories according to the blogger's interests, enabling the user to browse the information and opinions from bloggers concerned about the domains that they choose. The information about the blogger is cached for future use.

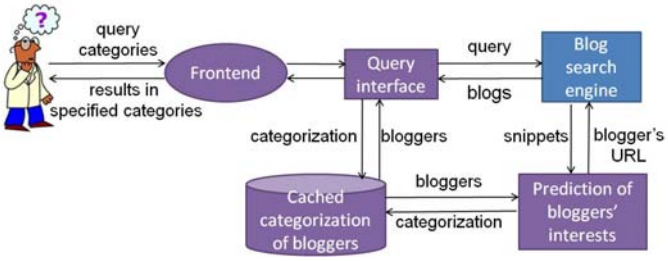

Figure 4 Architecture of Spectrum

Identifying Domain Experts

The main challenge in Spectrum is identifying bloggers with interests and expertise in a topic domain. One possible source of domain categorization of bloggers is blog directories. However, blog directories do not enable users to search for posts published by the bloggers they list. Moreover, the blogosphere is constantly changing, with bloggers joining and leaving. As a result, blog directories do not contain the most up-to-date information.

Instead of directly utilizing the blog directories, Spectrum learns a classification model for important bloggers in a particular domain, using the blog directories as training data. With this classification model, the system is then able to dynamically determine whether a blogger is worth reading given the domain selected by the user.

The blog posts published by bloggers provide important clues for predicting their interests. In addition, other blog posts citing the blogger indicate the importance of the blogger in the domain. If a blogger is an important author in a domain, most of his/her posts should be related to this domain, and he/she should also be consistently cited in the context of this domain. Therefore, the system utilizes the posts written by bloggers and the posts citing those bloggers to predict their interests and expertise of bloggers.

Instead of using the full content of the posts, Spectrum uses short snippets, consisting of the title and the first few sentences of a blog post. Using snippets eliminates the need to download complete web pages. Snippets are also faster to analyze than full text, enabling real time processing, which is especially critical for web applications.

There are two challenges in predicting bloggers' interests with blog snippets. First, blog articles are written in an informal erratic style. Bloggers sometimes even invent new words and grammars to express themselves idiosyncratically (Qu *et al.*, 2006). Second, bloggers do not confine themselves to one topic (Pew, 2006). A school teacher may blog about her personal life in addition to curriculum plans. Therefore, the post snippets by a blogger compose a multi-topic and noisy text corpus that is difficult to classify.

Categorizing Snippets of Blog Posts

To address these challenges, we propose a two-layer classification approach to predict bloggers' interests. For each blogger *b*, the system collects a set of recent blog snippets written by *b*, denoted as $P_b = \{p_0, p_1, ..., p_n\}$. A snippet consists of the title and the first two or three sentence from a post, containing about 40 words. The system also collects a set of snippets by other bloggers that have hyperlinks to the blog sites of *b*, denoted as $L_b = \{L_0, L_1, ..., L_n\}$. For a given domain category *c*, the task is to predict whether blogger *b* is an important author in that domain.

The proposed technique addresses this task with a two-layer classification model. In the first layer, the classifiers produce a probability estimate $p(c \mid s)$ for each snippet *s*, which is the probability that the snippet belongs to category *c*. The snippet could be a post snippet from *P* or a citation snippet from *L*. In the second layer, the system derives a set of features consisting of the categorization probabilities of the post snippets in *P* and the citation snippets in *L* respectively. The two sets of features are used together to predict the interests of *b*.

The first layer classifiers categorize the snippets. For a domain category c, we train a binary text classifier to estimate $p(c \mid s)$, the probability that a snippet s belongs to that category.

To build text classifiers of snippets, we take the content words of the snippets as features. We remove stop words (e.g., articles, pronouns, conjunctions, etc.) in snippets. The rest of words are stemmed. For each category, we selected the most predictive 2000 stemmed words according to Information Gain (Yang and Pedersen, 1997). To categorize the snippets, we use the Support Vector Machine (SVM) algorithm (Vapnik, 2000), which has been shown to be efficient and effective for text classification (Dumais *et al.*, 1998; Joachims, 1998). In our work, we use the sequential minimal model (SMO) developed by Platt (1998) to efficiently train the SVM classifier.

A standard SVM classifier makes binary predictions about the membership of instances *x* according to $y = sign(f(x))$, where $f(x)$ is the raw output of SVM. However, $f(x)$ is not a proper probability estimate of

$p(y \mid x)$. We utilize the method proposed by Platt (1999) to derive the probability of prediction by fitting the output of the SVM to a sigmoid model. The probability of membership is computed as follows:

$$p(y \mid x) = \frac{1}{1 + \exp(Af(x) + B)} \quad (1)$$

Here A and B are maximum likelihood estimates based on the training set $(y, f(x))$.

Encoding a Blogger

Before categorizing bloggers, we must describe how they are encoded. Categorizations of a blogger's post snippets and citation snippets provide important clues about a blogger's interests. The question is how to derive features for the blogger that can be used to predict the overall interests of the blogger.

For each category c, we take all the probability estimates $p(c \mid p_i)$ for $p_i \in P$. The set of probability estimates is

$$E(c) = \{p(c \mid p_0),..., p(c \mid p_i),...p(c \mid p_n)\} \quad (2)$$

$E(c)$ shows how much a blogger writes about category c. The probability estimates in the $E(c)$ are binned and placed into a histogram. For example, we sampled 30 snippets for each blogger in our experiment. For the category of law, a binary classifier was trained to classify law snippets. Using the classifier, we get a set of probabilities, $E(law) = \{p(law \mid p_0),..., p(law \mid p_i),...p(law \mid p_n)\}$ for the 30 snippets. Figure 5 shows the histogram for the set of probability estimates.

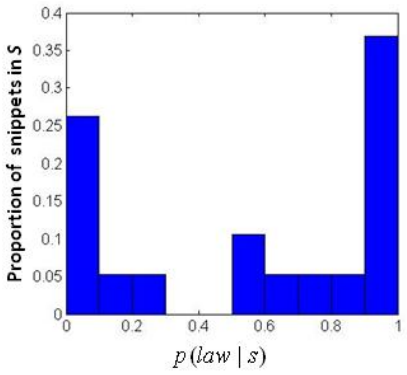

Figure 5 Distribution from a real sample of $E(law) = \{p(law \mid p_0),..., p(law \mid p_i),...p(law \mid p_n)\}$

We divide the [0, 1] range into K intervals and compute the proportion of snippets in P with $p(c \mid p_i)$ falling in each interval. This results in a K-element distribution for the category c. We use d_k^p to denote the kth element in the distribution for the category. It is the proportion of snippets in P with $p(c \mid p_i)$ falling in the kth interval. Formally, d_k^p is computed by Equation (3)

$$d_k^p = \frac{|P_k|}{|P|}, \text{where } P_k = \left\{ p \mid p \in P, p(c \mid s) \in [\frac{k-1}{K}, \frac{k}{K}) \right\} \quad (3)$$

We also calculate the mean and variance of $p(c \mid p_i)$. These are denoted and. $d^{p}{}_{main}$ and $d^{p}{}_{\text{var}}$. The proportions, mean and variance form a group of features D(P),

$$D(p) = \{d^{p}{}_{0},..., d^{p}{}_{K}, d^{p}{}_{main}, d^{p}{}_{\text{var}}\} \quad (4)$$

D(P) characterizes how much blogger b writes about the domain category. Similarly, the system derives another group of features D(L) from the categorization of citation snippets. D(L) characterizes how often blogger b is cited in the context related to the domain category. The two groups of features are combined together to characterize the interests of bloggers in category c.

$$D(c) = D(p) \cup D(L) \quad (5)$$

As identified in the studies of (Pew, 2006), bloggers may be interested in multiple domains. They may also write about topics not related to their main interests, such as personal stories and recent news. Furthermore, the categories are not independent from each other. For example, a law professor who writes about legal topics may also write a lot about political news. However, there is less chance that political posts would appear in an artist's blog that discusses oil paintings. Therefore, to capture the relation between topic domains, we use all the features derived for all of the categories to categorize blogger's interests. A blogger b is encoded as the union of $D(c_j)$ for $C = \{c_1, c_2,..., c_m\}$, as shown in Equation (6)

$$b = \{D(c_0), D(c_1),..., D(c_m)\} \quad (6)$$

Categorizing Bloggers' Interests

To categorize bloggers' interests, we train the second layer of classifiers using the derived features shown above. We experimented with a number of machine learning algorithms, including SVMs (Platt, 1999), nearest neighbor (Martin, 1995), and neural network with one hidden layer. An SVM with a linear kernel learns the weights of features and constructs a hyperplane to separate the positive and negative samples; these learned weights are helpful for explaining the trained classifier. Nearest neighbor and two-layer neural networks are able to model non-linear relationship between features. Specifically, the hidden layer in neural network allows the representation of sub-combination of features. Our experiment shows that the SVM achieves the highest precision and the neural network achieves the highest recall. Nearest neighbor performs the worst among the three classification methods. We describe details about the experiment in the next section.

Experiments

Dataset and Experiment Setup

Many blog directories have been created on the web to organize information and help users browse different topics in the blogosphere, for example, BlogCatalog and the blog section of Yahoo directory. The blog directories are compiled by expert editors or online communities. Within the directories, blog sites are organized into different topics. For our experiments, we collected lists of blog sites for eight major categories: art, business, education, health, law, politics, religion and technology. In our experiment, we assume that each blog site is owned by a single blogger. Although some blog sites are maintained by multiple people, they share similar interests. Altogether

we collected 4,428 bloggers for the 8 categories. We labeled each blogger with the categories assigned to their blog sites in the blog directories.

To collect blog snippets for the bloggers, we used Google Blog Search (2008). We queried the blog search engine with the URL of each blog site and collected the top 30 results for each blogger. The title and the search result summary returned by the search engine were used together as the snippet. Altogether we collected 86,598 blog post snippets for the 4,428 bloggers, resulting in 19.6 snippets per blogger on average. Because of the multi-topic nature of blogs, 130 bloggers with 2,689 snippets are categorized into multiple domains in the directories, which consist of 2.9% of the bloggers and 3.1% of the snippets in our collection.

We implemented the two-layer classification model described earlier using the Weka package (Witten and Frank, 2005), a Java-based knowledge learning and analysis environment developed at the University of Waikato in New Zealand.

In our experiment on the proposed two-layer classification model, we needed two separate datasets for classifies in each layer. We randomly divided the bloggers into two sets. The snippets retrieved for the first set 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 set of bloggers using 10-fold cross-validation.

To evaluate the classifiers in each layer, we used the conventional precision, recall and F1 measures. To evaluate the performance over all the categories, we computed the micro-averaged values for the three measures, which combine the performance of individual categories, weighted by the number of instances in the categories.

Categorization of Blog Post Snippets

To categorize the snippets of blog posts, we need labeled posts for training and testing the classifiers. However, this domain information for blog posts is not readily available. Although some blogs have tags, the tags are not semantically consistent and cannot be used reliably as labels. In our experiment, we propagated the domain of blogger's interests to their posts. Thus, the snippets of blog posts in our dataset were labeled by the interests of corresponding bloggers, which necessarily introduced some noise. According to the experimental setup described in the previous subsection, the classifiers were trained on the snippets of the first set of bloggers and tested on the snippets of the second set of bloggers. Specifically, there were 43,351 training snippets and 43,247 test snippets.

Recall that categorization of snippets is modeled with the "one-vs-all" scheme for multi-label classification. Binary classifiers were trained to distinguish the target category from the other categories. In our experiments, we applied an SVM with a linear kernel in the Weka package with default options. The micro-level F1 over all the categories is 0.526. Categorization of short snippets is a difficult task (Dumais and Chenn, 2000), so we did not expect to have very high accuracy. In our two-layer classification model, the results of snippet categorization are used to generate features for categorizing bloggers' interests, which is our ultimate goal. As shown in the following subsection, 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.

Categorization of Bloggers' Interests

We experimented with three different methods for the second-layer classification to predict a blogger's interest, SVM, neural network and nearest neighbor, all implemented in the Weka package. The SVM classifier uses linear kernel and default options in Weka. The Neural network classifier consists of one hidden layer with 8 nodes. All classifiers were tested on the second half of the dataset, which contains 2,214 bloggers with 43,247 snippets. We evaluate the performance with 10-fold cross-validation. Table 1 shows the performance of the three classification method based on 10-fold cross validation.

Table 1 Performance of SVM, neural network and nearest neighbor for categorizing bloggers' interests

	Precision	Recall	F1 measure
SVM	0.889	0.826	0.856
Neural Network	0.884	0.845	0.865
Nearest neighbor	0.830	0.813	0.821

As shown in Table 1, the performance of the SVM and neural network are comparable in terms of micro-F1. The Neural network achieves higher recall than the SVM, whereas the SVM achieves slightly higher precision than neural network. Nearest neighbor performs the worst in all three measures.

To evaluate overall performance in categorizing bloggers' interests, we compared the two-layer classification model with a baseline algorithm which categorized bloggers' interests directly from the text that they wrote. All the text snippets sampled for a blogger were mixed together to form a large text document. The linear form of the SVM was used to classify the mixed text documents and the results of text classification were directly used as predictions for the corresponding bloggers. This baseline was tested on the whole dataset using 10-fold cross-validation. Micro-level precision, recall and F1 measure were computed for the baseline algorithm.

Table 2 Comparison of the proposed method with the baseline

	Precision	Recall	F1 measure
proposed method	0.884	0.845	0.865
baseline	0.618	0.745	0.672
Improvement	40.7%	10.3%	25.7%

Table 2 compares the performance of the proposed two-layer model (using an SVM in the first layer, and a neural network in the second) with the baseline algorithms in term of the micro-level precision, recall and F1 measure. It shows that the mixture of blogger snippets is too noisy to

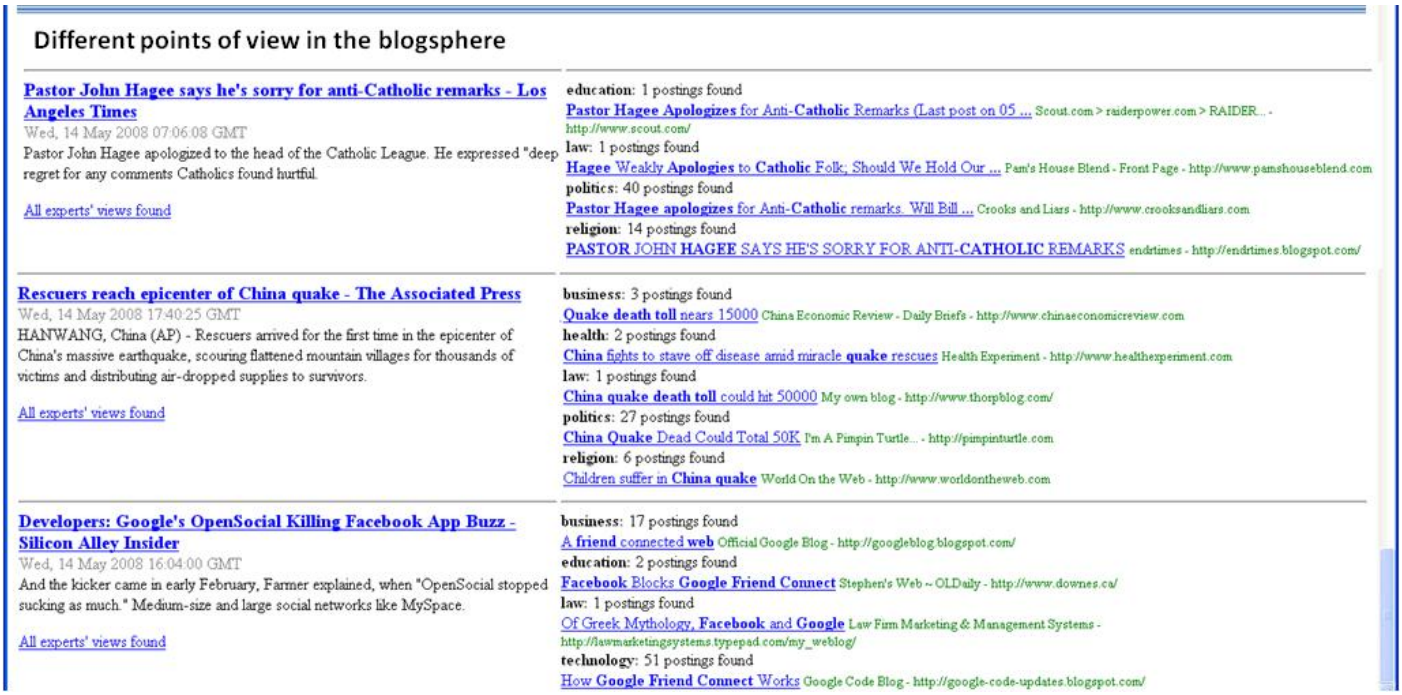

Figure 6 Aggregate different points of view about current

be accurately categorized by the baseline method. However, the two layer model is able to achieve high accuracy despite the errors in the first layer classifications shown in the previous subsection.

Multiple Perspectives about Current News

There is an ecological relationship between blogs and news media. Blogs are an important medium for general internet users to express opinions about current news events and topics. Pundits in the blogosphere in particular publish updates and analyses about news issues in their professional domains. The information and comments posted on blogs attract attention not only from individual news readers, but also from journalists, corporations, and government organizations. Nowadays it is not uncommon for journalists to cite comments and information from blogs. Businesses and governments view blogs as a valuable source for understanding opinions of the general public about news issues. To leverage this ecological relationship between blogs and news, we applied our model of multi-perspective blog search to the news context. The system retrieves a daily RSS feed for most popular news from Google News (2008). For each news issue, the system automatically aggregates blog posts related to that issue and categorizes them according to bloggers' interests. The system enables users to track opinions about current news and gain an understanding of the perspectives of bloggers with different interests and concerns.

There are two main steps to aggregating multiple perspectives around news issues. First, the system analyzes the retrieved news web page to extract a set of keywords for the news issue, using the method we developed in (Liu *et al.* 2007). Second, the keywords are used as queries to search for related blog posts in different categories via the multi-perspective blog search system. During the querying process, the system automatically selects all the categories and returns the categories with any search results.

Figure 6 shows a screenshot of a web page aggregating multiple perspectives for top news stories. Along with each news item, the system presents the number of blog posts it found for each category in the collected results. The aggregated blogs provide social context for news reading: what kinds of people are concerned about this issue, and what do they think about it. For the news items shown in the screenshot, political bloggers wrote extensively about the earthquake in China, whereas the news about Google's OpenSocial attracts attention from business people and technology enthusiasts. If users are interested in certain aspects, they can expand the list to view more posts from bloggers who are also concerned with that aspect. The posts provide additional details and opinions about the news issue from that particular perspective.

Conclusion

In this paper, we present Spectrum, a meta-search system for blogs that enables users to search for different points of view in the blogosphere. The system filters and categorizes blog search results according to the interests and expertise of the corresponding bloggers. We also describe multi-perspective blog search in the context of news-reading to

retrieve information and opinions around current news from multiple perspectives.

To predict bloggers' interests in Spectrum, we developed a two-layer classification model that categorizes bloggers' interests based on short snippets of posts written by the blogger and posts citing them. In the first layer, we predict the probability that a single snippets belongs to a domain. In the second layer, we derive two sets of features from the two sets of probabilities, one set from the post snippets and one set from the citation snippets. The derived features are then used to predict the bloggers' interests. Although short and noisy blog post snippets are hard to classify, the two-layer classification model was shown to be robust to the noise inherent in classifying individual snippets. We conducted experiments on a collection of bloggers compiled from blog directories, with blog post snippets retrieved from Google Blog Search. The proposed model achieves precision of 88.4% and recall of 84.5% in categorizing blogger's interests, outperforming the baseline algorithm (precision of 61.8% and 74.5%) which directly classifies the mixture of blogger's snippets.

References

Bhagat, S., Cormode, G. and Rozenbaum, I. 2007. "Applying link-based classification to label blogs". In Proceedings of WebKDD/SNAKDD 2007: KDD Workshop on Web Mining and Social Network Analysis.

Cointet, JP., Faure, E. and Roth, C. 2007. "Intertemporal topic correlations in online media". In Proceedings of the International Conference on Weblogs and Social Media.

Dumais, S. T., Platt, J., Heckerman, D. and Sahami, M. 1998. "Inductive learning algorithms and representations for text categorization". In Proceedings of 7th International Conference on Information and Knowledge Management.

Dumais, S., Chen, H. 2000. "Hierarchical classification of Web content" In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval

Durant, K. T. and Smith, M. D. 2006. "Mining Sentiment Classification from Political Web Logs". In Proceedings of Workshop on Web Mining and Web Usage Analysis at 12th ACM SIGKDD (WebKDD-2006).

Efron, M. 2004. "The liberal media and right-wing conspiracies: using cocitation information to estimate political orientation in web documents". In Proceedings of the 13th ACM international conference on Information and knowledge management.

Gamon, M., Basu, S., Belenko, D, Fisher, D., Hurst, M. and König, A. C. 2008. "BLEWS: Using Blogs to Provide Context for News Articles". In Proceedings of the International Conference on Weblogs and Social Media.

Google Blog Search http://blogsearch.google.com/. 2008

Google News. http://news.google.com/. 2008

Hu, J., Zeng, H.-J., Li, H., Niu, C., and Chen, Z. 2007 "Demographic prediction based on user's browsing behavior". In Proceedings of 16th International World Wide Web Conference.

Ikeda, D., Fujuki, T. and Okumur, M. 2006. "Automatically linking news articles to blog entries". In AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Joachims, T. 1998. "Text categorization with support vector machines: Learning with many relevant features". In Proceedings of European Conference on Machine Learning.

Klamma, R., Cao, Y. and Spaniol, M. 2007. "Watching the Blogosphere: Knowledge Sharing in the Web 2.0". In Proceedings of International Conference on Weblogs and Social Media (ICWSM'07)

Liu, J., Birnbaum, L. and Wagner E. 2007. "Compare&Contrast: Using the Web to Discover Comparable Cases for News Stories". In Proceedings of the 16th International Conference on World Wide Web

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

McCallum, A., Corrada-Emanuel, A., and Wang, X. 2005. "Topic and role discovery in social networks". In Proceedings of International Joint Conference of Artificial Intelligence.

Gilad Mishne and Maarten de Rijke. A Study of Blog Search. In Proceedings of ECIR-2006. LNCS vol 3936. Springer 2006.

Ni , X., Xue, G-R., Ling, X., Yu, Y. and Yang, Q. 2007. "Exploring in the weblog space by detecting informative and affective articles," In Proceedings of 16th International World Wide Web Conference.

Oberlander, J. and Nowson, S. 2006. "Whose thumb is it anyway? Classifying author personality from weblog text". In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics.

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

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,

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.

Qu, H. Pietra, A. L. and Poon, S. 2006. "Classifying blogs using NLP: Challenges and pitfalls". In AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs.

Rifkin, R. and Klautau, 2004. "A, In Defense of One-Vs-All Classification". The Journal of Machine Learning Research.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. 2006. "Effects of age and gender on blogging". In AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Steyvers, M., Smyth, P., Rosen-Zvi, Michal. and Griffiths, T. 2004. "Probabilistic author-topic models for information discovery". In Proceedings of the 10th international conference on Knowledge discovery and data mining.

Vapnik, V.N. 2000. "The Nature of Statistical Learning Theory". Springer-Verlag, New York, NY.

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

Yang, Y., Pedersen J.P. 1997. "A Comparative Study on Feature Selection in Text Categorization". In Proceedings of the 14th International Conference on Machine Learning.