Predicting Author Blog Channels with High Value Future Posts for Monitoring

Shanchan Wu
University of Maryland, College Park
wsc@cs.umd.edu

Tamer Elsayed
King Abdullah University of Science and Technology (KAUST)
tamer.elsayedaly@kaust.edu.sa

William Rand
University of Maryland, College Park
wrand@umd.edu

Louiqa Raschid
University of Maryland, College Park
louiqa@umiacs.umd.edu

Abstract

The phenomenal growth of social media, both in scale and importance, has created a unique opportunity to track information diffusion and the spread of influence, but can also make efficient tracking difficult. Given data streams representing blog posts on multiple blog channels and a *focal query post* on some topic of interest, our objective is to predict which of those channels are most likely to contain a *future* post that is relevant, or similar, to the focal query post. We denote this task as the *future author prediction problem* (*FAPP*). This problem has applications in information diffusion for brand monitoring and blog channel personalization and recommendation. We develop prediction methods inspired by (naïve) information retrieval approaches that use historical posts in the blog channel for prediction. We also train a ranking support vector machine (SVM) to solve the problem. We evaluate our methods on an extensive social media dataset; despite the difficulty of the task, all methods perform reasonably well. Results show that ranking SVM prediction can exploit blog channel and diffusion characteristics to improve prediction accuracy. Moreover, it is surprisingly good for prediction in emerging topics and identifying inconsistent authors.

Introduction

Social media is playing an ever increasing role in the marketing of new products and brands; this is in part because word-of-mouth communication, such as social media, have a dramatic effect on consumers' purchase decisions (Chevalier and Mayzlin 2006). Brand managers must pay attention to social media so that they can monitor the pulse of conversations that concern their brand (Li and Bernoff 2008). They can identify emerging discussions and join the conversations, possibly to encourage positive word-of-mouth (Godes and Mayzlin 2009).

Prioritizing or personalizing blogs or other social media channels is essential since managers do not have time to monitor the entire blogosphere. It is also useful to determine how quickly posts on a focal topic will spread across the blogosphere, and more importantly, which bloggers will post on that focal topic in the near future.

As an illustration, consider the Gap logo fiasco in the Fall of 2010.[1] Gap introduced a new logo, changing the iconic logo it had for 20 years almost overnight. There was an immediate outpouring of negative comments about the new logo on Twitter, Facebook, and across the blogosphere; Gap quickly reverted to the old logo. It would have been very helpful if a brand manager at Gap could have detected a blog post on this topic early on, and then predicted whether or not that conversation would spread to other blogs, and which bloggers, if any, would write about the topic. If the brand manager had this information, then she could select which blogs to monitor. She could participate in conversations, or even contact the bloggers ahead of time, to provide more accurate information and to keep them up to speed on the company's response.

To achieve that, it is necessary to develop tools that identify which blog channel is likely to next discuss Topic X (e.g., Gap Logo Redesign) as it relates to a brand (e.g., Gap), or which bloggers will respond to Topic X. To formalize these questions, we pose the following problem: *Given a focal query post on some topic on a blog channel, what other blog channels are likely to post on that topic in the (near) future?* The term *query post* refers to a post that will be used for search and for comparison[2]. We denote this task as the *Future Author Prediction Problem* (*FAPP*).

A good solution to the problem must predict the content of future posts to determine if they will be relevant to the query post. Then, for the relevant posts, one must predict the author blog channels. Finally, the joint expectation for these two prediction tasks must be maximized and the Top K authors/channels must be chosen. We note that predicting the content of a future post is difficult since there are few features that can be used for prediction. On the other hand, predicting the author of a future post is somewhat easier since we can consider the historical posts in a blog channel to build a profile of the author.

In this paper, we consider several solutions to *FAPP*. **PROF** and **VOTE** are inspired by information retrieval approaches and exploit historical posts to make a prediction. We also identify a number of additional features to train a

[1] "Gap to Scrap New Logo, Return to Old Design", Advertising Age, http://adage.com/article?article˙id=146417, October 11, 2010

[2] A similar term, *query document*, was used by Yang et al (Yang et al. 2009) to refer to a document whose phrases are used as queries.

ranking support vector machine for prediction, denoted as **RSVMP**. We test our methods using a blog dataset from Spinn3r (Burton, Java, and Soboroff 2009). Despite the difficulty of the *FAPP* task, all methods provide reasonably accurate results. **PROF** dominates **VOTE** while **RSVMP** dominates both. We also identify multiple characteristics that impact prediction accuracy including diffusion stage (*cRatio*), volume versus author count (*V/AC*) and blog channel consistency. **RSVMP** can exploit all of these characteristics to improve prediction accuracy.

These characteristics are of great interest since they affect the strategy and efficacy of a brand manager. For instance, if the topic is in the middle of its diffusion across the blogosphere (i.e., a mid-range cRatio), such as halfway through the Gap Logo controversy, then that is a critical period when the brand manager can have the greatest impact on the conversation. Before that time, it may not be clear if the topic will take off, and after that point, the conversation around it slows down, or perhaps has already trended negative. If the brand manager can predict which authors are likely to post in the mid-stage of diffusion, then actions can be taken. Our results show that **RSVMP** achieves accurate predictions under this scenario. It also performs surprisingly well for emerging topics.

Alternately, suppose that the story is not spreading, but is heavily-discussed only by a few authors (i.e, a high V/AC). If these authors are vocal (e.g., have a lot of followers), then it is important to predict new authors; this is another scenario where **RSVMP** can make accurate predictions.

Content-based techniques such as **PROF** are good at predicting the "usual suspects", however, what really concerns a brand manager is when a *difficult-to-predict* blogger or community gets involved. Difficulty increases when bloggers are inconsistent in their posts or because the comments come from a diverse set of bloggers. For instance, in the Gap Logo scenario, the brand manager may typically monitor clothing and fashion blogs, but the controversy may have emerged around blogs of graphic artists. While highly-consistent bloggers are easier to predict, **RSVMP** also performs well in identifying bloggers who are less consistent or have a diversity of profiles.

To summarize, we define a novel and challenging prediction problem *FAPP*. We develop multiple prediction methods and complete an extensive experimental evaluation. We show that a ranking SVM can be trained to exploit relevant features and can make accurate and useful predictions for many brand monitoring scenarios.

Related Work

Blog channel tracking and online news monitoring have become topics of research interest recently. For instance, the dynamics of the news cycle has been studied through the tracking of topics and memes (represented by soundbites) as they disseminate and evolve over time (Leskovec, Backstrom, and Kleinberg 2009). On the blog side, blogTrust (Varlamis, Vassalos, and Palaios 2008) examined the sudden convergence of communities of bloggers and their connection to real world events, while El-Arini et. al. (El-Arini et al. 2009) provided efficient techniques to sample posts in the blogosphere for personalized coverage and ranking. Since most of this work focused on tracking information as it spreads across communication channels, our high value blog channel prediction can complement this work by prioritizing which channels to monitor to achieve a better use of scarce resources.

Our work could also be beneficial even when the goal is a full catalog of all blogs. For instance, BlogScope (Bansal and Koudas 2007) has been very successful at online analysis of high volumes of blog channels; at present it indexes over 39 million blog channels and almost a trillion posts and updates the indexes every three hours (Bansal and Koudas 2007). Continuously updating an inverted index, can incur significant overhead, and so our blog channel prediction could provide a significant benefit by prioritizing updates to the index, based on user interests.

Problem Characteristics

Problem Definition

We define a blog channel as an event stream of posts or blog entries originating from a single source. The problem of predicting high value blog channels for monitoring is defined as follows:

Definition 1 ***Future Author Prediction Problem (FAPP):*** *Given a query post q posted at time T_q, identify the high value blog channels $B_{q,\Delta T}$ that will contain at least one future post p in the interval $(T_q, T_q+\Delta T]$ that is topically similar to q. Consider a similarity metric M_{sim} and a threshold η. Then $B_{q,\Delta T}$ must satisfy the following condition:*

$$\forall b \in B_{q,\Delta T}\ \exists\, p \in b \mid T_p \in (T_q, T_q+\Delta T] \wedge M_{sim}(p,q) \geq \eta$$

The goal is to identify up to K channels in $B_{q,\Delta T}$.

We decompose The *FAPP* into two sub-tasks: the *relevance task (RT)* is to identify unknown future posts p such that $M_{sim}(p,q) \geq \eta$, and the *authoring task (AT)* is to predict the blog channel B_p in which such a post p appears. *FAPP* is more complex and different from a traditional retrieval problem. For retrieval, the collection of all posts $B_{j,\Delta T}$, for all blog channels j, is known a priori. In contrast, for *FAPP* each future post p and its features are not known. A solution to *FAPP* must maximize the *joint expectation* for both tasks for post p with respect to query q and blog channel B_p, i.e., that the post p is relevant to the post q, and that B_p is the authoring blog channel for p.

Since *FAPP* is novel and difficult, in order to understand the quality of the results, we will perform an evaluation of the simpler *AT* for a *given* post, i.e., its features are known. While *AT* prediction is simpler, obtaining accurate results may be difficult since there is exactly one authoring blog channel for each post. In comparison, for *FAPP*, there may be many authoring blog channels in the ground truth.

Computing Similarity of Posts

We use the similarity between two posts as a proxy to indicate that the two posts are on the same topic. Posts are represented in the vector space model as a vector of terms where each term is weighted. In our experiments, we used the Okapi BM25 weighting function (Robertson et al. 1994)

The similarity between 2 vectors (posts) $\vec{V}_i$ and $\vec{V}_j$ is as follows:

$$M_{sim}(\vec{V}_i, \vec{V}_j) = \sum_t w^i_{tf}(t) \times w^j_{tf}(t) \times w_{idf}(t)$$

where $w^x_{tf}(t)$ is the *tf*-weight of term t in vector $\vec{V}_x$ and $w_{idf}(t)$ is the *idf*-weight.

Blog Channel Features

We consider several features related to the blog channel. The first is the **consistency**; we note that a channel that *consistently* posts on a specific topic indicates a kind of authority on that topic. To represent the consistency of a blog channel b, we use the average of the pairwise similarity scores between different historical posts of that channel. Formally, the consistency score $\psi(b)$ can be computed as follows:

$$\psi(b) = \frac{2}{m \cdot (m-1)} \sum_{p_i, p_j \in b, i \neq j} M_{sim}(\vec{V}_{p_i}, \vec{V}_{p_j})$$

where p_i and p_j are historical posts of blog channel b, and m is the number of historical posts in blog channel b.

We consider additional features including named-entities, links between channels, and links to external pages; they are discussed in the evaluation section.

Diffusion-Related Features

cRatio While there are many mathematical models of diffusion (Bass 1969), we propose a simple metric $cRatio$ to characterize the diffusion stage of a topic at time T. Consider a query post p in blog channel b at time T. Let $N_{history}$ and N_{future} be the number of blog channels other than b with posts that are similar to p before and after T respectively. We define $cRatio$ as follows:

$$cRatio = N_{history}/(N_{history} + N_{future})$$

V/AC The number of distinct authors posting on a topic is important for monitoring. We define a metric of blog volume versus author count, denoted as V/AC, as follows. For a query post p, $V/AC = N_{post}/N_{author}$ where N_{post} is the number of posts topically-similar to p and authored by N_{author} distinct blog channels during time period T to $T + \Delta T$.

Prediction Methods

Profile Based Prediction (PROF)

A profile of a blog channel can represent the content of its posts and should be updated as new posts appear. Maintaining profiles has been explored in several studies, e.g., (Roitman, Carmel, and Yom-Tov 2008); the key issues include the number of terms to maintain and the frequency at which the profile is updated. In this work, we adopt a temporal decay model which, for simplicity, does not consider absolute time; instead, the time interval between updates is used as a time unit.

Suppose $\{p_1, p_2, ..., p_n\}$ is a sequence of posts in blog channel b and each post p_i is represented as a weighted term vector $\vec{V}_{p_i}$. The blog channel profile vector $\vec{V}^1_b$ is initially set to $\vec{V}_{p_1}$ upon arrival of post p_1. As each new post p_i arrives at time unit i, the blog channel profile vector $\vec{V}^{i-1}_b$ is updated to $\vec{V}^i_b$ as follows:

$$\vec{V}^i_b = \theta \cdot \vec{V}^{i-1}_b + (1-\theta) \cdot \vec{V}_{p_i}$$

θ is a temporal decay factor, $0 < \theta < 1$; we choose an appropriate value for θ based on tuning from experiment datasets.

The profile based prediction algorithm is to retrieve the Top K blog channels ranked by the their similarity scores to a focal query post. The similarity of the profile of channel b to query post q is $Sim(q, b)$ and it is computed as follows:

$$Sim(q, b) = M_{sim}(\vec{V}_q, \vec{V}^n_b)$$

Voting-Based Prediction (VOTE)

VOTE chooses the top K channels using the aggregate similarity score of all historical posts in a channel b with a given query post q, which is computed as follows:

$$score(q, b) = \sum_{p_i \in b} M_{sim}(\vec{V}_q, \vec{V}_{p_i})$$

In computing the above score, **VOTE** considers only the Y ($> K$) most similar posts to q.

Ranking SVM Based Prediction (RSVMP)

Ranking SVM We represent the match of a blog channel to a query post as a vector $\vec{x}$. Each element in the vector is a numerical value indicating some correlation between the blog channel and the query post. There are different types of correlation between a blog channel and a query post and hence there are multiple elements in a vector $\vec{x}$. Any pair of vectors $(\vec{x}_i, \vec{x}_j) \in$ a ranking R if $\vec{x}_i$ ranks higher than $\vec{x}_j$ in R. Suppose that there is some optimal ranking R^* representing the ground truth. The goal is to find a ranking function f that approximates the optimal ranking R^*. A ranking function f is evaluated by comparing its ranking R^f with R^*.

In practice, R^* is not available. The ranking SVM is provided with training data corresponding to one or more partial rankings (partial orders) $R' \in R^*$. It can then learn a ranking function f from these partial orders. Consider a generic mapping from the feature vector $\vec{x}$ in the original feature space to a new feature vector $\phi(\vec{x})$ in a virtual feature space. Assume f is a ranking function as follows:

$$\forall(\vec{x}_i, \vec{x}_j) \in R' : f(\vec{x}_i) > f(\vec{x}_j) \Longleftrightarrow \vec{w} \cdot \phi(\vec{x}_i) > \vec{w} \cdot \phi(\vec{x}_j) \quad (1)$$

The goal is to learn an f which is concordant with the given partial orders $R' \in R^*$ and which can also generalize well beyond R'. A ranking SVM will obtain an approximate solution by solving the following optimization problem (Herbrich, Graepel, and Obermayer 2000):

$$minimize: \qquad \frac{1}{2}|\vec{w}|^2 + C \sum \xi_{i,j}. \quad (2)$$

subject to:

$$\forall(\vec{x_i}, \vec{x_j}) : \xi_{i,j} \geq 0 \quad (3)$$

$$\forall(\vec{x_i}, \vec{x_j}) \in R' : \vec{w}(\phi(\vec{x_i}) - \phi(\vec{x_j})) > 1 - \xi_{i,j} \quad (4)$$

$\xi_{i,j}$ are non-negative slack variables to allow some training error. C is a parameter that controls the trading-off between the margin size and training error, The solution weight vector w^* can be written in the form of training pairs. The ranking SVM will then use w^* for prediction, to rank the set of candidate blog channels, for some incoming query post.

Training Pairs Select N_{train} training query posts with posting time near T_{train}, where $T_{train} \leq T_c - \Delta T$. Get the ground truth of each training query post in the time range $(T_{train}, T_{train} + \Delta T)$. For each training query post, retrieve the top K' $(K' \geq K)$ blog channels, for each SVM feature, prior to T_{train}. All of these blog channels are candidate blog channels of that training query post. For a training query post, a candidate blog channel which is in the ground truth is set to be ranked higher than a candidate blog channel which is not in the ground truth. Each pair of them is composed to be a training pair.

Experimental Evaluation

Evaluation Datasets and Metrics

Dataset The dataset provided by Spinn3r.com is a set of 44 million blog posts crawled between August 1st and October 1st, 2008. We focus on blog channels with human authors rather than machine generated posts. We selected the posts that were published between July 30 and October 1 2008, the interval of interest. We then filtered out the blog channels that have less than 30 posts or more than 120 posts in the interval of interest. The statistics of the dataset that was used for the evaluation is in Table 1.

Query Posts and Ground Truth We created 2 sets of query posts, Q_1 and Q_2; these query posts are obtained from the beginning of an interval starting on September 1. We created three test datasets to obtain ground truth posts for Q_1 and Q_2. One test dataset included 2 days of posts from September 1 to September 2, another included 10 days of posts from September 1 to September 10, and a third included a 30 day dataset from September 1 to 30. A ground truth blog channel is one that includes at least one future post (in some test dataset) that is similar to the query post (in Q_1 or Q_2). We used an Okapi similarity score of 130 as the threshold to identify ground truth posts.

For example, Q_1 contains 861 query posts. Each focal post matched an average of 22 ground truth blog channels in the 2-day test dataset and 47 in the 10-day test dataset. The query posts in Q_2 had similar numbers of ground truth blog channels.

We used Amazon's Mechanical Turk marketplace to validate the Okapi metric by comparing its results with human judgement. We randomly selected 50 query posts. For each query post, we selected 3 candidate posts with a high Okapi score in the range [120, 800] and 3 candidate posts with a low Okapi score in the range [40, 60]. For each candidate post, we asked three users to evaluate the similarity between the candidate post and the query post. For the candidate posts with a high Okapi score, 93% of them were ranked as "very similar" or "similar" or "may be similar" by at least 2 users. This percentage was 17% for posts with a low Okapi score. Assuming "very similar", "similar", "maybe similar" to be one agreement, and "not similar" to be another agreement, then the inter-annotator agreement was 91% for posts with high scores and 83% for those with low scores.

Table 1: Statistics of the blog channel experiment data set

Time range	07/30/08–10/1/08
Number of blog posts	2,185,810
Number of blog channels	42,005
Avg number of posts per blog channel	52.04

Subset of Query Posts and Ground Truth

- Subset of query posts in Q_1 and Q_2 with different ranges of $cRatio$ values.
- Subset of consistent blog channels with consistency scores in the range of $[60, +\infty]$.
- Subset of query posts having V/AC in the range $[1.5, +\infty)$. The ground truth for this high V/AC was calculated in the 10-day test dataset.

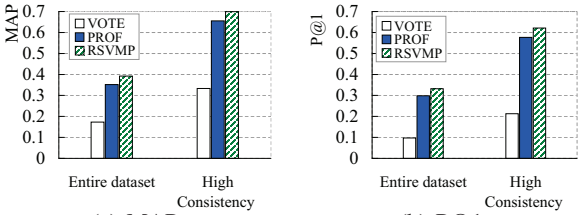
Training Data The training data was obtained from July 30 to August 31. All historical posts in this period, for each blog channel, were used by both **VOTE** and **PROF** in a straightforward manner. For **RSVMP**, we selected the training query posts at the start of an interval on August 22. We created 2 ground truth datasets. The first used posts that occurred within 2 days after August 22, and the second used posts that occurred within 10 days. The features of the blog channels that were used to produce the training pairs for each training query post were collected in the interval from July 30 to August 21. We reiterate that there was *no overlap* between the training data and testing data.

Feature Selection for RSVMP We consider the following features of a candidate blog channel b corresponding to a focal query post q:

- *FT-CHANNEL*: The similarity score between q and the profile of blog channel b.
- *FT-POST*: The similarity score between q and the post in b which is most similar to q.
- *FT-NE*: The similarity score between named entities extracted from q and the profile of b.
- *FT-PROFILE*: The similarity score between the profile of the author blog channel for q and the profile of b.
- *FT-CONSISTENCY*: The consistency score for b.
- *FT-OFFLINKS*: The (weighted) count of external links that are common to the author blog channel for q and b.
- *FT-INSIDELINKS*: The (weighted) count of links between the author blog channel for q and b.

Metrics and Parameters Mean Average Precision (MAP) is widely used for evaluating ranking methods (Manning, Raghavan, and Schutze 2008). We use the Wilcoxon signed-rank test (Wilcoxon 1945) to determine statistical significance. We also report on $P@1$, the precision of the Top 1 prediction, for the authoring task.

We set the temporal decay factor $\theta = 0.8$ for building blog channel profiles by tuning the training data. The K value of top K was set to 1000.



(a) MAP (b) P@1

Figure 1: The performance for the *AT* task.

Experimental Results

Baseline Results for the *AT* Task Figure 1(a) reports on the MAP for the 3 methods on two test datasets. The label *"Entire Dataset"* corresponds to focal query posts from Q_1; the ground truth blog channels is from the 10-day test dataset. The label *"High Consistency"* corresponds to Q_2 and the 10-day test dataset. These blog channels were filtered to only include consistent blog channels.

For the "Entire Dataset", *RSVMP* has a *MAP* value of 0.39, and *PROF* has an MAP of 0.35. Given that there is only one ground truth author for any post, these MAP values are surprisingly good, reflecting an accurate prediction. For the "High Consistency" channels, all methods show increased accuracy as expected. *MAP* is as high as 0.70 for *RSVMP*. This suggests that our methods perform with good accuracy on the *AT* task. We note that only 4 of the 7 correlation features were useful for this prediction; they are *FT-CHANNEL*, *FT-POST*, *FT-NE* and *FT-CONSISTENCY*.

Since there is only one author per post, Figure 1(b) reports on P@1, for all 3 prediction methods for the 2 datasets. As expected, these values are not as high as MAP. Nevertheless, they reflect a reasonable quality of prediction.

The Wilcoxon signed-rank test on MAP shows that *RSVMP* significantly outperforms *PROF* and *PROF* significantly outperforms *VOTE*, all with p far smaller than 0.01. We further note that while *RSVMP* can benefit from the training data, the improved accuracy of supervised learning over a naive *PROF* is limited.

Baseline Results for the *FAPP* Task Figure 2 reports on the MAP for the *FAPP* Task for the 3 methods. The test datasets labeled as *"Entire Dataset"* and *"High Consistency"* are the same as was used for the *AT* task. Unlike the *AT* task, where all 3 methods had reasonable prediction accuracy and where *PROF* and *RSVMP* showed very good performance, the *FAPP* task is much more challenging. For the "Entire Dataset", *RSVMP* has the best *MAP* value of 0.23 while *PROF* has a value of 0.20. For "High Consistency", the *MAP* increases to a value of 0.42 for *RSVMP*. We note that while these MAP values may appear to be low, they are comparable to the MAP values reported for the TREC blog distillation task (Ounis, Macdonald, and Soboroff 2008); there the reported MAP values are also in the range of 0.10–0.30. The Wilcoxon signed-rank test shows that *RSVMP* significantly outperforms *PROF* and *PROF* significantly outperforms *VOTE*, all with p far smaller than 0.01.

Table 2 reports on the MAP of all the methods, for focal query posts in Q_1, w.r.t. the test datasets of different time spans, for the FAPP task. Prediction accuracy for the 10-day test dataset is higher. This is probably because of the greater number of ground truth blog channels.

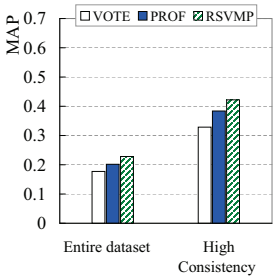


Figure 2: The performance for the *FAPP* task.

Table 2: MAP for the *FAPP* task w.r.t. different test datasets.

	VOTE	PROF	RSVMP
2-day test dataset	0.1383	0.1474	0.1672
10-day test dataset	0.1773	0.2018	0.2281

Table 3: The impact of *cRatio* on the "Entire Dataset" for the *FAPP* task.

cRatio	0-0.2	0.2-0.4	0.4-0.5	0.5-0.6	0.6-0.8	0.8-1.0
VOTE	0.090	0.167	0.234	0.222	0.137	0.062
PROF	0.144	0.193	0.257	0.244	0.151	0.056
RSVMP	0.188	0.225	0.288	0.262	0.170	0.070

Table 4: The impact of *cRatio* on the "High Consistency" test dataset for the *FAPP* task .

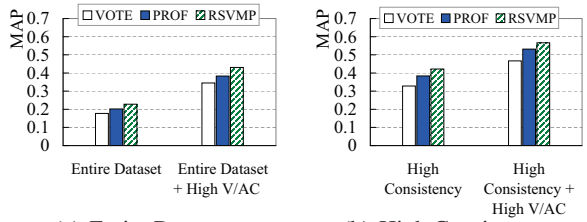
cRatio	0-0.2	0.2-0.4	0.4-0.5	0.5-0.6	0.6-0.8	0.8-1.0
VOTE	0.091	0.233	0.498	0.437	0.228	0.110
PROF	0.172	0.285	0.578	0.487	0.262	0.214
RSVMP	0.205	0.309	0.605	0.525	0.314	0.235

Impact of Diffusion Stage (cRatio Values) Recall that $cRatio = N_{history}/(N_{history} + N_{future})$. N_{future} is the number of blog channels other than b with similar posts after T in the 30 day test dataset. $N_{history}$ is the number of blog channels other than b with similar posts before T in the 30 day training dataset.

Table 3 reports on the MAP values for the *FAPP* task, for the 3 methods, for the focal test query posts from Q_1. The ground truth is from the 10-day test dataset. The results are grouped by the $cRatio$ values for the query posts. Table 4 reports on the MAP for the same methods for the focal query posts from Q_2. The ground truth is from the consistent blog channels in the 10-day test dataset.

PROF outperforms *VOTE* and *RSVMP* dominates both. The value of MAP is highest for all the methods when $cRatio$ is in the range 0.4–0.5 and 0.5–0.6, i.e., the middle stage of diffusion. When $cRatio$ is in the range 0.4–0.5, and for consistent blog channels, *RSVMP* has an MAP value that is as high as 0.61. *RSVMP* also does surprisingly well for emerging topics.

Impact of Blog Volume Versus Author Count Figure 3(a) reports on the MAP for the 3 methods for the *FAPP* task. The left part reports on the *"Entire Dataset"* and the right reports on *"Entire Dataset"* with high values for V/AC. Figure 3(b) reports on the MAP for consistent blog channels. The left part reports on the *"High Consistency"* and the right reports on *"High Consistency"* with high values for V/AC.



(a) Entire Dataset (b) High Consistency

Figure 3: The impact of V/AC for the *FAPP* task.

With increased values of V/AC, prediction accuracy improves, across all methods and across all datasets. This is consistent since high V/AC reflects repeated posts by some authors, thus making the *FAPP* prediction task somewhat easier. Further, *RSVMP* has an MAP value of 0.57 in Figure 3(b) for the *"High Consistency"* test dataset with high values for V/AC. This reflects that we can predict repeated posts on the same topic by consistent authors, with high prediction accuracy, or high confidence in the prediction.

Difficult and Diverse Predictions We note from the previous discussion that *RSVMP* is able to exploit multiple features and provide a more accurate prediction even in a difficult prediction scenario corresponding to an emerging topic. A further analysis of the properties of the predicted blog channels of *PROF* and *RSVMP* illustrate that the predictions of *RSVMP* may have high utility. For example, we compared the consistency scores of the Top 10 predictions for the two methods over the entire dataset. The average score is 108.6 for *RSVMP* and 112.7 for *PROF*. We also compared the average profile similarity scores between the predicted blog channels and the focal query post over the entire dataset. The average score is 399 for *RSVMP* and 428 for *PROF*. Thus, *RSVMP* was able to successfully identify the *less consistent* authors who have not posted on the focal topic in the past but who will post on the topic in the future. Similarly, *RSVMP* was able to successfully identify authors whose profile was not similar to the focal query post, but who nevertheless authored a post that was similar to the focal post. To summarize, these less consistent authors or authors with dissimilar profiles who nevertheless will post on the focal topic in the future may have more utility for the task of monitoring.

Conclusions

In this paper we proposed the *Future Author Prediction Problem*, and three potential solutions, *VOTE*, *PROF* and *RSVMP*. The most sophisticated method *RSVMP* significantly outperforms the others, but the straightforward method *PROF* still performs well. We also found that consistency, diffusion stage ($cRatio$), and blog volume versus author count (V/AC) all impact prediction accuracy. Prediction accuracy increases for consistent blog channels, and with regards to diffusion stage, prediction accuracy is better in the emerging stage than in the declining stage and highest in the middle stage. Prediction accuracy also increases when V/AC is high for the query posts. Although $cRatio$ and V/AC themselves may contain future information, estimates of their current values could be inferred from the historical data; we leave this examination for future work. In the situation where diffusion and blog factors can not be controlled, they can still be used to indicate a confidence level for the prediction accuracy of a given query post and provide additional information for recommendation.

Acknowledgements: This research was partially supported by NSF awards CMMI 0753124, IIS 0960963, and IIS 1018361.

References

Bansal, N., and Koudas, N. 2007. Blogscope: A system for online analysis of high volum text streams. In *Proceedings of the International Conference on Very Large data Bases (VLDB)*.

Bass, F. M. 1969. A new product growth for model consumer durables. *Management Science* 15(5):215–227.

Burton, K.; Java, A.; and Soboroff, I. 2009. The icwsm 2009 spinn3r dataset. In *Proceedings of the Conference on Weblogs and Social Media (ICWSM 2009)*.

Chevalier, J., and Mayzlin, D. 2006. The effect of word of mouth online: online book reviews. *Journal of Marketing Research* 43:348–354.

El-Arini, K.; Veda, G.; Shahaf, D.; and Guestrin, C. 2009. Turning down the noise in the blogosphere. In *KDD '09*.

Godes, D., and Mayzlin, D. 2009. Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science* 28(4):721–739.

Herbrich, R.; Graepel, T.; and Obermayer, K. 2000. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, 115–132. Cambridge, MA: MIT Press.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *KDD '09*.

Li, C., and Bernoff, J. 2008. Groundswell: Winning in a world transformed by social technologies. *Harvard Business School Press*.

Manning, C.; Raghavan, P.; and Schutze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Ounis, I.; Macdonald, C.; and Soboroff, I. 2008. Overview of the trec-2008 blog track. In *Proceedings of the Text REtrieval Conference (TREC)*.

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.; and Gatford, M. 1994. Okapi at TREC-3. In *TREC-3*, 109–126.

Roitman, H.; Carmel, D.; and Yom-Tov, E. 2008. Maintaining dynamic channel profiles on the web. *Proc. VLDB Endow.* 1(1):151–162.

Varlamis, I.; Vassalos, V.; and Palaios, A. 2008. Monitoring the evolution of interests in the blogosphere. In *IEEE ICDE Workshops*, 513–518.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. In *Biometrics Bulletin*, volume 1, 80–83.

Yang, Y.; Bansal, N.; Dakka, W.; Iperiotis, P.; Koudas, N.; and Papadias, D. 2009. Query by document. In *WSDM '09*.