Future Link Prediction in the Blogosphere for Recommendation

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

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

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

Abstract

The phenomenal growth in both scale and importance of social media such as blogs, micro-blogs and user-generated content, has created a need for tools that monitor information diffusion and make recommendations within these platforms. An essential element of social media, particularly blogs, is the hyperlink graph that connects various pieces of content. There are two types of links within the blogosphere; one from blog post to blog post, and another from blog post to blog channel (an event stream of blog posts). These links can be viewed as a proxy for the flow of information between blog channels and to reflect influence. Given this assumption about links, the ability to predict future links can facilitate the monitoring of information diffusion, making recommendations, and word-of-mouth (WOM) marketing. We propose different methods for link predictions and we evaluate these methods on an extensive blog dataset.

Introduction

The amount of new content and links being generated within social media, including blogs, micro-blogs and user-generated content, is increasing dramatically. When creating their posts, bloggers use hyperlinks to refer to pages and websites including the posts of other bloggers and more often, not just one particular post, but another entire blog channel. The collection of all blogs, i.e., the blogosphere, can be viewed as a dynamically changing representation of content streams, with an overlay of links. These links, in turn, can be viewed as a proxy to indicate the direction of information flow and influence in the blogosphere (Gruhl et al. 2004; Kempe, Kleinberg, and Tardos 2003; Song et al. 2007).

On the user side, a consumer of blogs might see a stream of interesting posts, and wonder if there are other blog channels that will link to the focal blog channel in the future. Such a link could indicate that the other blog channel is interested in a similar topic. Moreover, for a creator of a post on a focal channel, it is important to inspire a conversation around a particular topic, and so they want to know who will link to them. The timeframe of prediction is an additional critical factor given the stream of content in the blogosphere. Recently published posts typically attract more readers than older posts. However, recent posts may not have had sufficient time exposure to attract many links. Hence, accurate future link prediction is an important element when making recommendations for recent posts.

Future link prediction can also be a key element of a successful word-of-mouth (WOM) marketing strategy, by allowing manager's to predict the course of future WOM. Since understanding the future state of the link structure of the blogosphere can help bloggers, consumers of blogs and marketing managers, in this paper, we address the problem of *future link prediction (FLP)*. Informally, FLP is as follows: Given a focal blog channel, predict the Top K blog channels that will contain at least one future post that will link to the focal blog channel or some post in it.

There are two approaches to link prediction that have been successfully applied in other contexts (Leroy, Cambazoglu, and Bonchi 2010; Liben-Nowell and Kleinberg 2003). One approach is content-based, i.e., we compare the content of the focal post to the content of all other blog channels and choose the one with the closest content match. The second approach is network-based and utilizes the structural properties of the focal blog channel and the other blog channels to infer missing structure. At times these two viewpoints seem at odds; network science has often suggested the dominance of the structure of a network, while content-based approaches rarely utilize network structure. Our view is that these approaches are not mutually exclusive but rather they are two ends of a spectrum. We believe that a hybrid of structural- and content-based properties is needed to make accurate predictions in the blogosphere.

To examine this hypothesis, we apply several topological metrics that use historical links for prediction, including *Jaccard*, *CommonNeighbors*, and *Bonacich*. To efficiently calculate *Bonacich*, we present a method *Bonacich-A* based on an approximate *Bonacich* score. Moreover, we incorporate an additional network metric; *CommonExternal* is a method based on common external links. Besides link features, we also explore content features. We examine a content based method *CBP*; it uses a temporal profile to represent the interests and content of a blog channel based on historical posts. In addition, we propose a simple unsupervised learning based hybrid solution *HYBRID* that combines the features of *Bonacich-A* (the best solution among the link-based prediction methods) and *CBP*. Finally, we present a supervised learning method *RSVMP*, which is a ranking support vector machine for FLP. We report on the results of an evaluation on a blog dataset from Spinn3r.

Related Work

Several approaches have been successfully applied to link prediction. One class of solutions focuses on topological features of graphs (Liben-Nowell and Kleinberg 2003; Song et al. 2009). A second class uses machine learning approaches such as spectral transformation (Kunegis and Lommatzsch 2009), Markov Random Field Model (Wang, Satuluri, and Parthasarathy 2007), collective classification (Taskar et al. 2003), etc. An excellent summary is presented in (Leroy, Cambazoglu, and Bonchi 2010).

Taskar et al. (Taskar et al. 2003) applied a collective classification approach to predict links in relational data and entity-relationship graphs. This approach works well for labeled graph datasets where there are strong relationships and/or the nodes have rich feature labels that can be uniformly applied. We do not expect such methods to perform well in the blogosphere since there are no strong relationship types nor are there uniform labels. Another limitation is that such classification approaches may not scale well to the large graphs typical of social media.

An array of methods for link prediction based on topological features were presented in Liben-Nowell and Kleinberg (Liben-Nowell and Kleinberg 2003) who evaluated them on co-authorship networks. We use some of these same methods, but on blog data, and also compare them with other methods based on additional features.

Problem Definition

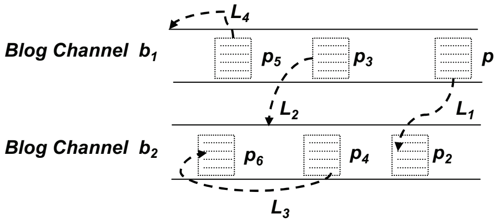

Figure 1: Links in the blogosphere

We define a blog channel as an event stream of posts (i.e., blog entries) originating from a single source (a blogger, organization, etc.). Figure 1 depicts two blog channels with three posts and links between them.

A link in a post that points to another post is a "post-to-post" link. An example of this type of link is L_1 in Figure 1, which points from post p_1 to p_2. A link that points to a blog channel is a "post-to-blog-channel" link; L_2 from post p_3 to blog channel b_2 is an example of this type of link. We abstract both types of links as "blog-channel-to-blog-channel" links. If a link points to the blog channel in which it appears, or to another post of the focal channel, then it is "self-referential"; L_3 and L_4 are examples of this type.

Definition 1 ***Future Link Prediction (FLP) Problem:*** *Given a focal blog channel b at a specific time point T and a time interval ΔT, identify K blog channels $B^b_{T,\Delta T}$ that will contain one (or more) future post(s) in the interval $(T, T + \Delta T]$ having at least one link pointing to blog channel b or any post of blog channel b.*

Prediction Methods

Link Based Prediction Methods

We apply the methods surveyed in (Liben-Nowell and Kleinberg 2003) to create structural metrics to the blogosphere, including **Jaccard** and **CommonNeighbors**. The metric Katz (Katz 1953) measures the status of a node by the total number of paths linking it to other nodes in the graph; an exponential discount is used as the path length increases. Bonacich (Bonacich 1987) generalized Katz's metric and proposed *Bonacich centrality*. It too reflects the total number of paths originating from a node and uses an attenuation factor α to discount indirect links and β to discount direct links. We propose a prediction method, **Bonacich-A**, based on an approximate *Bonacich* score. We use parameter D to approximate the computation. Only the paths having length up to D will be calculated. In general, lower values of D will have a significant impact on on the prediction output. For the sparse blog-channel-to-blog-channel link graph, D did not have much impact.

The link prediction methods **Jaccard**, **CommonNeighbors** and **Bonacich-A** are all based on links within the dataset or between different blog channels. Many links point to pages outside the blog dataset. We propose the method **CommonExternal** based on external links and inspired by the TF/IDF metric popular in information retrieval. For each external link URL_a that is common to blog channels i and j, suppose the link appears N_i (N_j) times in the corresponding blog channel i (j). Suppose that there are B_a blog channels that contain URL_a. Then, the weighted contribution to $score(i, j)$ by $URL_a = \frac{N_i \cdot N_j}{B_a}$. The final score $score(i, j)$ is the summation of all scores contributed by all common external links between the two channels.

Content Based Prediction Method (CBP)

CBP uses the blog channel profile and a similarity metric to make a prediction. We adopted a temporal decay model to update the profile. 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, 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 within experimental datasets.

After the profiles are built and indexed, **CBP** will retrieve the top *K* blog channels ranked by their profile similarity scores to the focal blog channel. We use a version of the state-of-the-art Okapi formula (Robertson, Walker, and Hancock-Beaulieu 1998) to calculate similarity.

Hybrid Prediction Method (HYBRID)

We consider a simple unsupervised learning approach that combines the predictions of *CBP* and *Bonacich-A* (the best predictor from the link based methods). HYBRID will be used as a baseline for comparison with a supervised learning approach. Recall that *CBP* and *Bonacich-A* both generate a Top K ranked list, and so the task becomes to merge these lists. There are many methods to merge ranked lists; a popular approach is based on the Borda count. While it is a simple solution it has the drawback that it gives equal weight to all rankings. In FLP, when there are historical links, then the prediction made by *Bonacich-A* is often superior to that made by *CBP*. We develop a method *HYBRID* that is inspired by the Borda count but favors the ranking of *Bonacich-A* when there are historical links; exact details of the method are available upon request.

Ranking SVM Based Prediction (RSVMP)

We apply ranking SVM (Joachims 2002) to rank the set of candidate blog channels for prediction, for a focal blog channel. We consider the following features for training the ranking SVM and we report on their effectiveness: (1)*FT-PROFILE*: The similarity score between the profile of the focal blog channel b and the profile of candidate blog channel b'; (2) *FT-INSIDELINKS-BONACICH*: The Bonacich score of the candidate blog channel b' with respect to the focal blog channel b using blog-channel-to-blog-channel links; (3) *FT-INSIDELINKS-COMMONNEIGHBOR*: The *CommonNeighbors* score of the candidate blog channel b' with respect to the focal blog channel b using blog-channel-to-blog-channel links; (4) *FT-EXTERNALLINKS*: The *CommonExternal* score of the candidate blog channel b' with respect to the focal blog channel b based on external links.

Experimental Evaluation

Evaluation Dataset and Metrics

Dataset We use a dataset provided by Spinn3r.com, which 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.

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
Average number of posts per blog channel	52.04
Number of external links (links pointing to outside of the dataset)	7,883,004
Number of bog-channel-to-blog-channel links without self references	154,218

Test Datasets and Training Data We created two test datasets to obtain ground truth. One test dataset included 10 days of posts from September 1 to September 10, another included 30 days of posts from September 1 to October 1. The subset of blog-channel-to-blog-channel links that are used to determine the ground truth are those links starting from a post in the test data, and pointing to a post in a focal blog channel or pointing to a focal blog channel. We exclude "self-referential" links. This created two sets of focal blog channels, S_1 and S_2. S_1 is the set of focal blog channels that contain ground truth in the 10-day test dataset, and S_2 is the set of all blog channels that contain ground truth in the 30-day test dataset. S_1 includes 3636 focal blog channels while S_2 includes 6831. The training data extends from July 30 to August 31.

Metrics and Parameters We report the values of Mean Average Precision (MAP). We selected $K = 20$ for evaluation since more than 90% of the focal blog channels have no more than 20 ground truth blog channels, i.e., channels which would later link to them. The values of α, β, θ were tuned from the training data, where $\alpha = 0.002$, $\beta = 1.0$, $\theta = 0.8$. We selected $D = 10$ for *Bonacich-A*. For this dataset, there was no benefit for values of D greater than 10, while there was a significant computational overhead.

Experimental Results

Baseline for the 3 methods Figure 2 reports on MAP for all of the methods on the two test datasets. CBP which utilizes content for prediction has the lowest prediction accuracy. All other methods which utilize historical links for prediction have higher accuracy than CBP. This shows that links are the most significant prediction feature for FLP. Among the link based methods, *Bonacich-A* dominates *Jaccard* and *CommonNeighbors* and *CommonExternal*. Recall that *CommonExternal* exploits external links for prediction. Its accuracy is similar to *CommonNeighbors* and this demonstrates that external links are also a good feature for prediction. *HYBRID* can benefit from combining links and content based features. Finally, the supervised learning method *RSVMP* dominates all of other methods.

For each prediction method, the MAP value for the 10-day test dataset is higher than the 30-day test dataset. This is a surprising and interesting result, since 10-days is a more narrow window for the methods to be correct. This reflects our argument that the blogosphere evolves in many ways. Both the topical interests of the bloggers and their continuing interest in following bloggers changes over time. For example, we observe that 14.6% of the focal blog channels in S_1 which have ground truth in the 10-day test dataset do not have historical blog-channel-to-blog-channel links in the training dataset. In comparison, 27.0% of the focal blog channels in S_2 which have ground truth in the 30-day test dataset do not have historical blog-channel-to-blog-channel links in the training dataset. This is consistent since the interval of the training data may be quite distant in time from some events, i.e., the posts that contain links in S_2. In other words, the significance of the historical links reduces over time as the interest of blog followers changes over time.

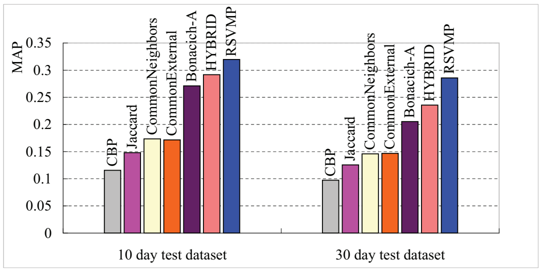

Figure 2: The performance of the prediction methods.

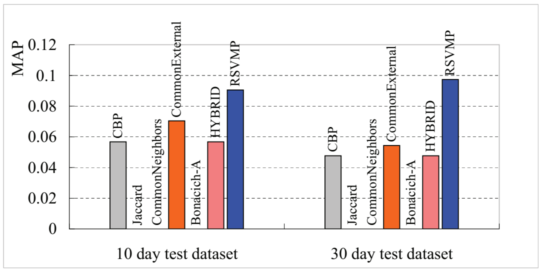

Figure 3: The performance of the prediction methods on the subset of the focal blog channels with 0 blog-channel-to-blog-channel historical links.

Subset with no historical blog-channel-to-blog-channel links Figure 3 reports MAP values for all of the methods on the subset of the focal blog channels without historical bog-channel-to-blog-channel links. The methods *Jaccard*, *CommonNeighbors* and *Bonacich-A* which only use bog-channel-to-blog-channel links for prediction can make no prediction and have a 0.0 MAP value. *HYBRID* has the same MAP value as *CBP*. *CommonExternal* outperforms *CBP*, indicating that the external links are more significant predictors compared to the blog channel profile alone. As expected, *RSVMP* dominates all methods.

Feature analysis for RSVMP Figure 4 reports MAP values for *RSVMP* method for different features. We consider two groups of features. One group of features is based on blog-channel-to-blog-channel links, i.e. network features. There are two features in this group: *FT-INSIDELINKS-BONACICH*, and *FT-INSIDELINKS-COMMONNEIGHBOR*. The other group of features are content based and include *FT-PROFILE* and *FT-EXTERNALLINKS*. Note that while *FT-EXTERNALLINKS* represents links, the value of these links are the content of the referenced pages. Figure 4 shows that *RSVMP* has better performance when applying the group of network based features alone, in comparison to applying the content based features alone. Also as expected, *RSVMP* has the best performance when it combines both groups of features.

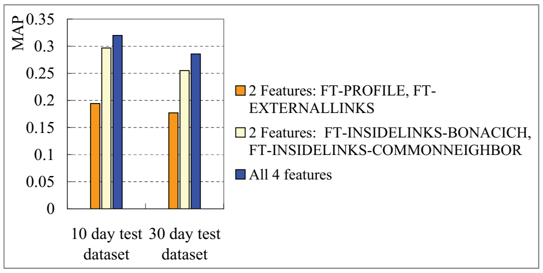

Figure 4: The performance of the *RSVMP* by applying different features.

Conclusions

In this paper, we address the problem of future link prediction in the blogosphere. Given a focal blog channel, we predict which blog channels will be most likely to contain at least one post having links to the focal blog channel or posts of the focal blog channel in the near future. We compare multiple link prediction methods, and show that a method which combines the network properties of the blog with content properties does better than methods which examine network properties or content properties in isolation. In future work, we will consider the joint influence of a group of blog channels. We will also try to predict the number of future links. This will allow us to identify what will be the most influential blog channels in the near future.

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

References

Bonacich, P. 1987. Power and centrality: A family of measures. *The American Journal of Sociology* 92(5):1170–1182.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW '04*.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *KDD '02*.

Katz, L. 1953. A new status index derived from sociometric analysis. *Psychometrika* 18:39–40.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *KDD '03*.

Kunegis, J., and Lommatzsch, A. 2009. Learning spectral graph transformations for link prediction. In *ICML '09*.

Leroy, V.; Cambazoglu, B. B.; and Bonchi, F. 2010. Cold start link prediction. In *KDD '10*.

Liben-Nowell, D., and Kleinberg, J. 2003. The link prediction problem for social networks. In *CIKM '03*.

Robertson, S.; Walker, S.; and Hancock-Beaulieu, M. 1998. Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. *TREC 1998* 199–210.

Song, X.; Chi, Y.; Hino, K.; and Tseng, B. L. 2007. Information flow modeling based on diffusion rate for prediction and ranking. In *WWW '07*.

Song, H. H.; Cho, T. W.; Dave, V.; Zhang, Y.; and Qiu, L. 2009. Scalable proximity estimation and link prediction in online social networks. In *IMC '09*.

Taskar, B.; Wong, M.-F.; Abbeel, P.; and Koller, D. 2003. Link prediction in relational data. In *NIPS '03*.

Wang, C.; Satuluri, V.; and Parthasarathy, S. 2007. Local probabilistic models for link prediction. In *ICDM '07*.