Connecting Mutually Influencing Bloggers

Aditya Pal, Jaya Kawale
Department of Computer Science
University of Minnesota
Minneapolis, MN 55455, USA
{apal,kawale}@cs.umn.edu

Abstract

The blogosphere shows the characteristics of a power law distribution where a small set of the bloggers (*influentials*) get the majority of readership and the vast majority receives little traffic. Blogger recommendation algorithms aim at finding *influentials* for recommendation, putting bloggers with limited readership at further disadvantage. These bloggers could benefit from mutual endorsement of each other with the eventual goal of forming strong local communities with broader readership. In this paper, we propose a recommendation algorithm to connect blogger pairs with the intent that once connected the bloggers would share a mutually influencing relationship between them. In particular, we compute bloggers' influence profile based on how much she influences her blog friends and recommend bloggers with similar influence profiles. We characterize bloggers into four different groups: {global leaders, connectors, local leaders, isolates}. Our result shows marginal benefit for isolates and significant benefit for local leaders. Our approach can be instructive in building intelligent recommendation engine for bloggers with limited readership to build strong local communities.

Introduction

The blogosphere shows the characteristics of a power law distribution where a small set of the bloggers (*influentials*) get the majority of readership and the vast majority receives little traffic (Shirky 2003). Blogger recommendation approaches that focus on recommending *influentials* do not capture the essential characteristics of blogging, i.e., it is a form of social interaction among people (Kumar et al. 2003). Kumar et al. distinguishes the blogosphere from the World Wide Web on two main characteristics: a) The blogosphere is replete with local community interactions amongst a small neighborhood of bloggers, and b) The time of the post is an important feature for a blog post but not so important for a web site.

In this paper, we propose an algorithm for recommending bloggers to each other. The recommendation algorithm aims at maximizing the influence two bloggers could extend on each other. The proposed benefit of this approach is that it could improve the mutual readership of the two bloggers and lead to bi-directional communication between the bloggers (*influential* and *non-influential* do not share this kind of relationship). We propose an influence estimation algorithm to estimate the influence of a blogger in her neighborhood. We use the influence estimates to find bloggers with similar influence profiles for recommendation to each other. We categorize bloggers based on their blogging rate and network characteristics in four categories: {global leaders, connectors, local leaders, isolates}. Our results indicate that by using our recommendation algorithm, local leaders benefit significantly in the improvement of their page rank and also have higher likelihood of having a mutually influencing relationship. Our work can be used by search engines and recommendation engines to find and connect similar bloggers leading to stronger bonding between the bloggers. The proposed algorithms are generic and can also be applied in other domains such as microblogs.

Related Work

The study of influence in the spread of topics and ideas in the blogosphere has been performed by (Gruhl et al. 2004; Gill 2004; Java et al. 2006). There is a rich foundation in the study of diffusion in networks with decades worth of work from various fields including sociology, biology, economics, epidemiology, physics and marketing (Wasserman and Faust 1994). (Gruhl et al. 2004) studied the dynamics of information propagation identifying chatter and spikes of topics. They use the Independent Cascade model of infectious diseases to model the propagation of topics and information. When a blogger writes on a topic there is a probability associated with his neighbors to read it and to further go ahead and write about it.

(Java et al. 2006) used the models proposed by (Kempe, Kleinberg, and va Tardos 2005) to maximize the spread of influence in a social network and found influential nodes. They apply the models to blogosphere to predict influential blogs. Their experiments also show that effective prediction of influential can be done with a simple heuristic based page rank technique.

(Tayebi, Hashemi, and Mohades 2007) describes a scheme for ranking blogs based on assigning two scores for each blog, called the personality score and the operational score. The scores capture the preferred bloggers of a blogger, or the vote from one blogger to another, with weights defining the probability of reaching blog B from blog A in a random surfer model. (K Fujimura 2005) describes a scheme

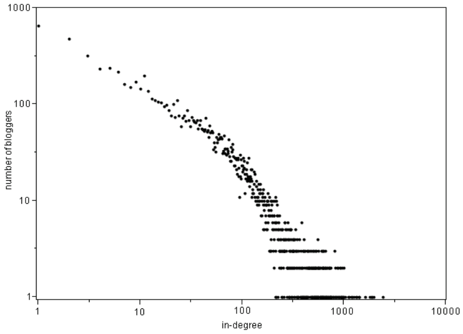

Figure 1: Log-Log plot of degree distribution of core nodes.

based on pagerank (Brin and Page 1998) to rank blogs and bloggers.

Our work differs from the prior work as it aims at recommending bloggers such that they share mutually influencing relationship in future. We present an influence estimation algorithm for connected bloggers that considers the blogs posts as causal links where one influences the other and measure the actual influence of a post on another using document similarity measures over temporally ordered posts. The proposed methodology could improve the bonding between bloggers in the long tail and lead to the formation of strong communities.

Dataset

We crawled LiveJournal blog graph by choosing 5 popular bloggers[1] as seed. The resulting **directed** graph contained 528,029 nodes, 2,113,724 edges and 2,000 core nodes. Core nodes are nodes whose neighbors and neighbors' neighbors are present in the crawled graph. Figure 1 shows the degree distribution of the core nodes. The plot resembles a power-law distribution with a skew towards the tail. For all the nodes, we collected their blogs during March 2009 - June 2009 (4 months). We collected 9,578,030 blogs with an average of 18 blogs per blogger.

Blogger Categorization

We categorized the core nodes into four different categories based on their blogging rate and graph prominence. We selected page rank (Brin and Page 1998) as a feature to measure the prominence of a given node in the blog graph. Page rank algorithm looks at how the nodes are interconnected and their relative importance to compute iteratively the page rank of each node. We calculated blogging rate as average number of blogs written by the user in a month. It gives us an estimate of how engaged is the person in the blogging.

We computed the mean of page rank and blogging rate for all the core nodes and partitioned them into four categories based on the two mean values. Figure 2 shows the scatter plot of users in the four categories. The four categories are defined as follows:

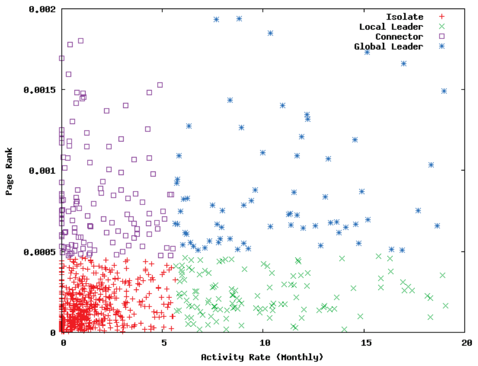

Figure 2: Scatter plot of bloggers in different categories (image is zoomed for better visualization).

- **Global Leaders**: Bloggers with high blogging rate and high page rank. These are the top influential bloggers who publish blogs frequently.
- **Connectors**: Bloggers with low blogging rate but high page rank. These bloggers are influentials yet they do not post as frequently as global leaders. Typically, celebrities and other famous people lie in this category. They do not blog very often yet enjoy a very high follower count.
- **Local Leaders**: They are bloggers who post frequently but are not as important as global leaders. They are bloggers who enjoy readership in their local community and are motivated towards publishing blogs.
- **Isolates**: Bloggers with low blogging rate and low page rank. Not many people connect with these bloggers and they are typically registered to follow other bloggers and post comments.

Our Algorithm

The blogosphere is represented as a directed weighted graph $G(V, E)$, where V represents the set of bloggers and E represents the set of directed edges between them. The directed edge $y \rightarrow x$ indicates "x reads y". A blogger x is considered to be reading blogger y's blogs if x friends y, comments or cites y. The weight I of the directed edge $y \rightarrow x$ indicates "influence of y on x" and is written as $I(y, x)$.

Mutually Influencing Relationship (MIR)

Mutually influencing relationship (MIR) is based upon mutual endorsements. The relationship between an influential and a non-influential blogger is non-mutual as the non-influential blogger reads the influential blogger's blogs, comments, cites or bookmarks it, but the influential blogger does not reciprocate the same. On the other hand there are various communities among the long tail of bloggers that thrive on mutual following of each others' blogs. MIR is a bi-directional relationship between two connected bloggers x,y, s.t. they influence each other approximately equally. The bloggers x and y must satisfy the constraint $M_\tau(x, y)$: $\frac{1}{\tau} \leq \frac{I(x,y)}{I(y,x)} \leq \tau$, where τ is the approximation threshold.

[1] http://en.wikipedia.org/wiki/List_of_LiveJournal_users

Algorithm 1 Influence Estimation Algorithm

```
Require: blog posts $P : \{p_1, p_2, .., p_n\}$, sorted on publish time in non-decreasing order. $T(p_i)$ indicates the time of the post. $DS(p_i, p_j)$ computes document similarity.
  for $i$ : 1 to $n$ do
    $u : user(p_i)$
    $U : u \cup in\text{-}neighbors(u)$
    $pl : latestpost(u)$   {latest post published by $u_i$ before $p_i$}

    {ps: post influence set}
    $ps : \{p \in P : (user(p) \in U) \wedge (T(pl) \leq T(p) < T(p_i))\}$

    {update influence estimates}
    for $v \in U$ do
      {$I(v, u)$ : 0, if not defined previously}
      $I(v, u) : I(v, u) + \frac{\sum_{p \in ps \wedge user(p)=v} [1 + DS(p, p_i)]}{\sum_{p \in ps} [1 + DS(p, p_i)]}$
    end for
  end for
  {normalize influence estimates}
  for $u \in all\text{-}users$ do
    $n : num\text{-}posts(u)$
    $U : u \cup in\text{-}neighbors(u)$
    for $v \in U$ do
      $I(v, u) : \frac{I(v,u)}{n}$   $\{\sum_{v \in U} I(v, u) = 1\}$
    end for
  end for
```

Influence Estimation Algorithm

We use the observation that blog posts are often the result of social interactions at the local level. A blogger draws influence from her blogger friends and vice-versa. The influence estimation algorithm presented by Algorithm 1 begins by constructing a post influence set (ps) which contains blog posts in u's neighborhood which could have influenced u to publish p_i. ps contains all the posts published by u's friends (including u) between two consecutive posts of u. For all the posts in ps we estimate the influence on p_i by computing their document similarly (DS). document similarity is computed by turning blog posts into tf-idf vectors and taking cosine similarity measure of the tf-idf vectors. Influences are then normalized with the constraint that each user draws a influence of 1 from her neighborhood.

Recommendation Algorithm

Recommendation algorithm (see Algorithm 2) aims to find bloggers that if connected would share a MIR between them. To do this, we first compute the user's influence profile (Ipr) containing three components: 1) Self-influence, 2) Influence of the user on her friends, 3) Ratio of her out degree with in-degree. Algorithm then uses the hypothesis that: *bloggers with similar influence vectors are more likely to have a* MIR, to find bloggers within same category that have closest influence profile (under τ approximation).

Results

Our algorithm has two parameters, τ, W. We experimented with several values of them and choose the ones that perform the best ($\tau = 1.5, W = [.2\ .5\ .3]^T$).

Algorithm 2 Recommendation Algorithm

```
Require: influence estimates $\{I(x, y) : \forall x, y \in V\}$. $\tau$ is approximation threshold for $MIR$. $cat(u)$ is one the four categories to which blogger $u$ belongs. $W$ is a vector with three components, s.t. the sum of the components is 1.
  for $u \in all\text{-}users$ do
    $U : in\text{-}neighbors(u)$
    $V : out\text{-}neighbors(u)$
    $out(u) : \sum_{v \in V} I(u, v)$
    $Ipr(u) : [I(u, u)\ \ out(u)\ \ \frac{|V|}{|U|}]$
  end for

  $R(u, v) : [Ipr(u) - Ipr(y)] \cdot W, \forall u, v \in V$
  for $u \in all\text{-}users$ do
    $U : in\text{-}neighbors(u)$
    $V : out\text{-}neighbors(u)$
    $y : \arg\min_v \{R(u, v) \leq \tau : \forall v \notin U \cup V \wedge cat(v) = cat(u)\}$
    {if $y$ exists, then create edges: $u \rightarrow y, y \rightarrow u$}
  end for
```

		I	LL	C	GL
Isolates	I	**.2**	.04	.05	.06
Local Leader	LL		**.42**	.02	.08
Connector	C			**.14**	.13
Global Leader	GL				**.17**

Table 1: Likelihood of $M_{1.5}$ satisfiability for two bloggers in various categories.

Likelihood of Mutually Influencing Relationship

We compute the influence estimates between users and estimate the probability of $M_{1.5}$ satisfiable relationships for core nodes (see Table 1). As expected, this probability is higher for bloggers in the same categories than across categories (one way t-test with $p < 0.001$) - justifying the claim that ties between an influential and a non-influential blogger are unidirectional with asymmetric flow of influence. The likelihood of MIR for LL users is highest indicating that they participate in mutual endorsements more than users of any other category.

Performance of Recommendation Model

One recommendation is generated per core node. If a node is already recommended then another recommendation including that node is discarded. Overall this leads to a generation of less than 1000 recommendation for the 2000 core nodes. To measure the model performance, we hold out 20% of the most recent blogs and use the remaining 80% of the blogs for influence estimation. Once recommendations are added to the graph, the hold out blogs (20%) are used to recompute estimation and test M_τ satisfiability of the edges added by the recommendation algorithm. Precision is calculated as the ratio of M_τ satisfiable recommendations by all the recommendations.

We choose two baseline algorithms: a) *Random recommendation* of users within the same category, b) *Document similarity* model, which computes average document similarity $DS(.,.)$ of all the blog post pairs of the two bloggers and picks blogger pairs with highest similarity for recom-

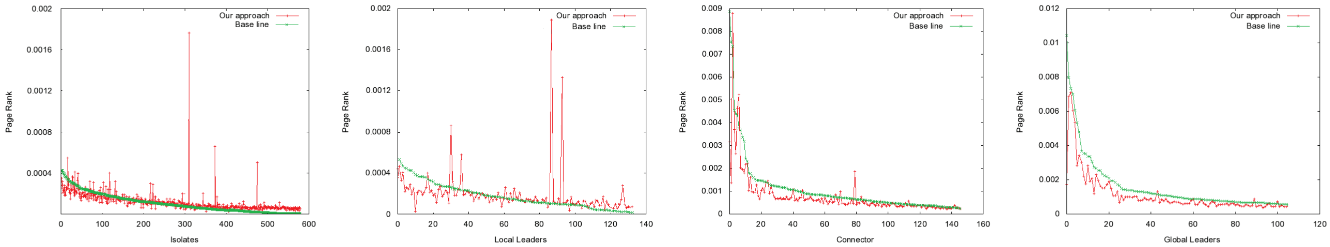

Figure 3: Category wise change in the page rank of bloggers after using our recommendation algorithm.

Category	I	LL	C	GL
Our recommendation model	.47	.73	.55	.56
Document Similarity	.32	.48	.35	.38
Random recommendations	.13	.37	.15	.15

Table 2: Precision of different models.

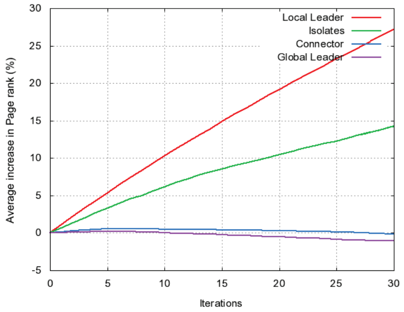

Figure 4: Change in Average Page Rank with our recommendation model compared to document similarity model.

mendation. Table 2 shows that our model improves over the best baseline models by 46-57%. This increase in performance is substantial. High precision for LL indicates that our model can find blogger pairs that would exhibit mutual relationship (indicator of high activity correlation and document similarity).

Measuring Changes in Page Rank

We measure the changes in page rank of nodes before and after the application of the recommendation algorithm to see how the recommendations affect nodes' graph prominence. The aim here is to see how the prominence of users would vary in the real world if they were to follow the recommendations proposed by our algorithm. The recommendation algorithms are run for 30 iterations and in each iteration the graph is updated as per the recommendations. Figure 3 shows that the page rank increases for bloggers in long tail if they are effectively connected and it could lead to decrease in the prominence of *influentials*.

Figure 4 shows that comparative change in page rank of our vs document similarity model . It shows that the increase obtained from using our models is more than 25% for LL and 14% for I for 20-25 iterations. This indicates that effective connections of the users between categories can lead to significant improvements in their page rank (used as an indicator of readership).

Conclusion

In this paper, we present a novel approach to recommend bloggers based on their influence profiles. We show that bloggers in the long tails could be effectively connected to create strong bonds between the connected users. We hypothesize that this could improve their readership and motivation to blog and formation of strong local communities.

Our work can be instructive for recommendation engines for providing effective recommendations to the users in blogosphere and other similar domains like microblogs. We aim to test our findings on microblog users in future work. It can also be useful for community managers to develop methods to retain people by recommending them to each other and enable the scope of mutual endorsements.

References

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Network ISDN Systems* 30(1-7):107–117.

Gill, K. E. 2004. How can we measure the influence of the blogosphere? In *WWW: workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW*, 491–501.

Java, A.; Kolari, P.; Finin, T.; and Oates, T. 2006. Modeling the spread of influence on the blogosphere. Technical report.

K Fujimura, T Inoue, M. S. 2005. The eigenrumor algorithm for ranking blogs. In *WWW, Workshop on the Weblogging Ecosystem*.

Kempe, D.; Kleinberg, J.; and va Tardos. 2005. Influential nodes in a diffusion model for social networks. In *ICALP*, 1127–1138.

Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2003. On the bursty evolution of blogspace. In *WWW*, 568–576.

Shirky, C. 2003. Power laws, weblogs, and inequality.

Tayebi, M. A.; Hashemi, M. S.; and Mohades, A. 2007. B2rank: An algorithm for ranking blogs based on behavioral features. In *WI*, 104–107.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.