

A Study of Communities and Influence in Blogosphere

Nitin Agarwal
Arizona State University
Tempe, AZ 85287, USA
Nitin.Agarwal.2@asu.edu

ABSTRACT

Blogging becomes a popular way for a Web user to publish information on the Web. Bloggers write blog posts, share likes and dislikes, voice opinions, provide suggestions, and report news. In this work we study influential bloggers in both community as well as individual blogs. We synthesize virtual communities by aggregating individual blogs with similar interests. We formulate the problem for identifying influential bloggers and synthesizing virtual communities from individual blogs, present a preliminary model, discuss the challenges, and pave the way for building a robust model that allows finding various types of the influentials. To illustrate these issues, we conduct experiments with data from a real-world blog site, evaluate multi-facets of the problem, and discuss unique challenges. We conclude with interesting findings and future work.

1. INTRODUCTION

The advent of participatory Web applications (Web 2.0 [16]) has turned the former mass information consumers to the present information producers [6]. Examples include blogs, wikis, social annotation and tagging, media sharing, and other such services. A “blog” is a weblog at a website where the entries by individuals are displayed in reverse chronological order. A typical blog can combine text, images, and links to other blogs and to Web pages. These entries can be blog posts or comments - the follow-up posts linked to some specific posts. Blogging is becoming a popular means for mass Web users to express, communicate, share, collaborate, debate, and reflect. Blogosphere is the virtual universe that contains all blogs. Bloggers, the blog writers, loosely form their special interest communities where they share thoughts, express opinions, debate ideas, and offer suggestions interactively. Blogosphere provides a conducive platform to build the communities of special interests. These multi-authored blog sites, also called community blogs are a good source of discussion on a product or an event. Others, single-authored blog sites often called individual blogs

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IDAR08 Proceedings of the Second SIGMOD PhD Workshop on Innovative Database Research (IDAR 2008), June 13, 2008, Vancouver, Canada.
Copyright 2008 ACM 978-1-60558-211-5/08/06 ...\$5.00.

are more personal journals or views. Some individual blogs have really good insights about events or products.

Some recent numbers from Technorati show a 100% increase in the size of Blogosphere every six months, “...about 1.6 Million postings per day, or about 18.6 posts per second”. Blogosphere has grown over 60 times during the past three years. Since new blog posts are being generated with such a blazing fast rate, novel ways have to be developed in order to keep track of everything happening in Blogosphere. Inspired by the high impact of the influentials in a physical community [9], we study a novel problem of identifying influential bloggers at community blogs. Nevertheless, the blogosphere consists of more individual blogs than community blogs. So we extend the study to include individual blogs as well. However, since individual blogs are single authored, it is insensible to find influential bloggers. This is achieved by synthesizing a virtual community of similar individual blogs. We aggregate the individual blogs that are similar and treat them as a virtual community. Next, we present the problem statement.

2. PROBLEM STATEMENT

As we discussed above, blogs can be *individual* and *community blogs*. For an individual blog, the host is the only one who initiates and leads the discussions and thus is naturally the influential blogger of his/her site. For a community blog where many have equal opportunities to participate, it is feasible to study who are the influentials. However, similar individual blogs could be aggregated to synthesize virtual communities and study the influential bloggers. We first present the problem statement of identifying influential bloggers in a community blog and then present the problem statement to aggregate similar individual blogs.

2.1 Identifying Influential Bloggers

Each blog post is often associated with some metadata like author of the post, annotations/tags, date and time of posting, and comments. In addition, one can also collect statistics like *outlinks* - posts or articles to which the author has referred, *inlinks* - other posts that refer to this post, post length, number of comments per post, and the rate at which comments are posted on a blog post.

In the simplest case, one can approximate an influential blogger with an active blogger who posts frequently. However, this is not the case in a physical world, where a voluble person is not necessarily or seldom influential, we are inquisitive whether we can employ the above metadata and statistics to identify influential bloggers. The search for in-

fluent bloggers boils down to the question on how to define an influential blogger.

An intuitive way of defining an influential blogger is to check if the blogger has any influential blog post, i.e., *A blogger can be influential if s/he has more than one influential blog post.* Assume we have an influence score for a post p_i , $I(p_i)$ For a blogger b_k who has N blog posts, $\{p_1, p_2, \dots, p_N\}$, their influence scores can be ranked in descending order, and her influence index, $iIndex(b_k)$ can be defined as $\max(I(p_i))$, where $1 \leq i \leq N$. Given a set U of M bloggers, $\{b_1, b_2, \dots, b_M\}$, the problem of identifying influential bloggers is defined as determining an ordered subset V of K bloggers, $\{b_{j_1}, b_{j_2}, \dots, b_{j_K}\}$ that are ordered according to their $iIndex$ such that $V \subseteq U$ and $K \leq M$, i.e. $iIndex(b_{j_1}) \geq iIndex(b_{j_2}) \geq \dots \geq iIndex(b_{j_K})$. V contains K most **influential bloggers**. For all the blog posts $\{p_1, p_2, \dots, p_L\}$ by all M bloggers, **influential blog posts** are those whose influence scores are greater than $iIndex(b_{j_K})$ or, $I(p_l) \geq iIndex(b_{j_K})$ for $1 \leq l \leq L$. Hence, we have the following corollary: those bloggers who published blog posts that satisfy $I(p_l) \geq iIndex(b_{j_K})$, for $1 \leq l \leq L$ will be called influential bloggers because their $iIndex$ will be greater than or equal to $iIndex(b_{j_K})$.

2.2 Aggregating Similar Individual Blogs

Since influential bloggers could be identified only within a community so we need to synthesize communities from individual blogs in order to include the invaluable and inordinate amount of data source of individual blogs. Bloggers, with their blogging behavior, tend to create social relationships with peer bloggers. However, most of them are locally connected with limited links to other bloggers, thus in the Long Tail [4]. The goal is to aggregate similar individual blogs. Given a blogger b , similar bloggers to b are a set of bloggers $B = b_1, b_2, \dots, b_n$, who share common patterns as b , like blogging on similar topics. Basically, every pair b, b_j of bloggers, where $1 \leq j \leq n$, blog on similar topics or share the latent process that inspires them to do so. Nevertheless, b, b_j may or may not have either linked to each other in their blog posts or present in the other's social network.

The problem of finding similar individual blogs can be formulated as: given a blogger b , identifying a set of bloggers B , such that every pair of bloggers b, b_j , where $1 \leq j \leq n$, satisfies the definition of similar bloggers. Similarity can be defined by topics, bag of words, tag clouds, etc., and will be discussed in Section 3.3.

3. PROPOSED MODEL

We first propose a preliminary model of identifying influential bloggers using some desirable properties of influence approximated by collectable statistics, then discuss some approaches to aggregate the similar individual blogs.

3.1 An initial set of intuitive properties

Following [9], one is influential if s/he is recognized by fellow citizens, can generate follow-up activities, has novel perspectives or ideas, and is often eloquent. Below we examine how this initial set of intuitive properties can be approximated by some collectable statistics.

1. **Recognition** - An influential blog post is recognized by many. This can be gauged by the inlinks (ι) to post p . The more influential the referring posts are, the more influential the referred post becomes.

2. **Activity Generation** - A blog post's capability of generating activity can be indirectly measured by how many comments it receives, the amount of discussion it initiates. Hence, a large number of comments (γ) indicates that the post can be influential. However, spam comments that do not add any value to the blog posts or blogger's influence could be eliminated using recent research [15].

3. **Novelty** - Novel ideas exert more influence as suggested in [9]. The number of outlinks is an indicator of a post's novelty. A large number of outlinks (θ) may suggest that a post refers to many other blog posts or articles, indicating that it is less likely to be novel.

4. **Eloquence** - An influential is often eloquent [9]. This property is most difficult to approximate using some statistics. Given the informal nature of the blogosphere, there is no incentive for a blogger to write a lengthy piece. Hence, a long post often suggests some necessity of doing so. Therefore, we use the length of a post (λ) as a heuristic to measure blog post's influence.

The above form an initial set of properties possessed by an influential post. It is evident that each of the above four may not be sufficient on its own, and they should be used jointly in identifying influential bloggers. Starting with this initial set, we next build a preliminary model that allows us to examine, analyze, modify, and extend the model.

3.2 Influence graph - a preliminary model

Blog-post influence can be visualized in terms of an influence graph or *i-graph* in which the influence of a blog post flows among the nodes. Each node of an i-graph represents a single blog post characterized by the four properties (or parameters): ι, θ, γ and λ . i-graph is a directed graph with ι and θ representing the incoming and outgoing influence flows of a node, respectively. Hence, if I denotes the influence of a node (or blog post p), then *InfluenceFlow* across that node is given by,

$$InfluenceFlow(p) = w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n) \quad (1)$$

where w_{in} and w_{out} are the weights that can be used to adjust the contribution of incoming and outgoing influence, respectively. p_m denotes all the blog posts that link to the blog post p , where $1 \leq m \leq |\iota|$; and p_n denotes all the blog posts that are referred by the blog post p , where $1 \leq n \leq |\theta|$. $|\iota|$ and $|\theta|$ are the total numbers of inlinks and outlinks of post p . From Eq. 1, it is clear that the more inlinks a blog post acquires the more recognized it is, hence the more influential it gets; and an excessive number of outlinks jeopardizes the novelty of a blog post which affects its influence.

The influence (I) of a blog post is also proportional to the number of comments (γ_p) submitted to that blog post,

$$I(p) \propto w_{com} \gamma_p + InfluenceFlow(p) \quad (2)$$

where w_{com} is the weight to regulate the contribution of the number of comments (γ_p) towards the influence of p .

We consider blog post quality as one of the parameters that may affect influence of the blog post. Although there are many measures that quantify the goodness of a blog post such as fluency, rhetoric skills, vocabulary usage, and blog content analysis, for the sake of simplicity, we here use the length of the blog post as a heuristic measure of the goodness of a blog post in the context of blogging. We define a weight

function, w , which rewards or penalizes the influence score of a blog post depending on the length (λ) of the post. The weight function could be replaced with appropriate content and literary analysis tools. Combining Eq. 1 and Eq. 2, the influence of a blog post, p , can thus be defined as,

$$I(p) = w(\lambda) \times (w_{com}\gamma_p + InfluenceFlow(p)) \quad (3)$$

The above equation gives an influence score to each blog post. Note that the four weights can take more complex forms and can be tuned.

According to the definition of influential blogger in Section 2, a blogger can be considered influential if s/he has at least one influential blog post. For a blogger B , we can calculate the influence score for each of B 's N posts and use the maximum influence score as the blogger's influence index or *iIndex*,

$$iIndex(B) = \max(I(p_i)) \quad (4)$$

where $1 \leq i \leq N$. The top k among the total bloggers are the most influential ones. Thresholding is another way to find influential bloggers. However, determining a proper threshold is crucial to the success of such a strategy and requires more research.

3.3 Aggregating Similar Individual Blogs

We now describe an approach to aggregate similar individual bloggers and synthesize a virtual community. We use data from a blog site directory available at BlogCatalog (<http://www.blogcatalog.com/>). BlogCatalog organizes the blog sites under pre-specified categories. Bloggers can specify the categories where a blog site should be categorized. The number of categories keeps increasing as more blog sites are submitted to BlogCatalog, although in a controlled fashion. At the time of writing, BlogCatalog had in total 56 level-1 categories. The maximum depth of the category structure is 3. Additional information collectable from BlogCatalog includes categories under which the blogger lists his/her blog site, blog post level tags, blog site level tags, and snippets of last 5 blog posts for each blog site.

We leverage the label information to cluster the blog sites. A naïve way could be to treat all the blog sites that have the same label as one cluster resulting in too many clusters and moreover some of the clusters thus obtained might be related and would be better if they are merged into one cluster. Also many blog sites are tagged under more than one labels, which makes it difficult to form clusters in the naïve way. To achieve this, we cluster similar labels.

Clustering the similar labels can be formulated as an optimization problem. Assume we have t labels, l_1, l_2, \dots, l_t and are clustered into k clusters, C_1, C_2, \dots, C_k , then optimal clustering is obtained if, for any two labels l_i and l_j ,

$$\min \sum d(l_i, l_j), \forall (l_i, l_j) \in C_m, 1 \leq m \leq k, i \neq j \quad (5)$$

$$\max \sum d(l_i, l_j), \forall l_i \in C_m, \forall l_j \in C_n, 1 \leq (m, n) \leq k, m \neq n \quad (6)$$

Here $d(l_i, l_j)$ refers to a distance metric between the labels l_i and l_j . (5) minimizes the within-cluster distance between the cluster members and (6) maximizes the between-cluster distance. Finding efficiently an optimal solution for the above min-max conditions is infeasible.

Other approaches like K-means requires the value of k *a priori*. Another approach to cluster the blog sites is based

on the blog post tags and blog site tags. Each blog site can be profiled based on these accumulated tags. A simple cosine similarity distance metric could be used to find similarity between different blog sites. However, the vector space model of the blog sites based on the tags is high-dimensional and sparse. We use a SVD based clustering algorithm as the baseline to avoid the curse of dimensionality. We chose top 25 eigenvectors to transform the blogs to reduced concept space. Pairwise similarity between blogs was computed using cosine similarity between reduced concept space vectors of the blogs. More details could be found in [1].

Based on the above discussion and limitations with the vector space model, we propose an approach to achieve blog site clustering leveraging the "collective wisdom" of the bloggers. Often bloggers specify category labels for a particular blog site. Such blog sites help in establishing links between these labels. This results in a *label relation graph*. For example, labels like **Computers** and **Technology**; **Computers** and **Internet**; **Computers** and **Blogging** were linked by the bloggers. The number of blog sites that create the links between various labels is termed as *link strength*, which could be treated as the edge weights of the label relation graph. Using this label relation graph, different labels can be clustered or merged. We call this link-based clustering, *WisClus*.

WisClus clustering approach is highly time sensitive and adaptive to blog dynamics, since labels of a blog site could change depending on what the blogger is blogging about. This results in dynamic as well as adaptive clustering. Every time new blog posts appear, there will be new edges appearing in label relation graph or the link strength changes as blogger specifies different labels, leading to changes in the clustering results. Since the blogosphere provides more emphasis on the freshness of the content, the proposed clustering approach would reflect similar dynamics.

4. CHALLENGES

Here we discuss the challenges while studying influential bloggers and aggregating the similar individual bloggers.

Identifying Influential Bloggers The preliminary model presents a palpable way of identifying influential bloggers. However, several issues remain to be addressed,

- Are active bloggers influential?
- How can we evaluate the model's performance in absence of typical training and test data. Hence, it requires innovative ways of evaluating and validating the end results.
- How do we handle the subjectivity aspect of the problem as different people may have disparate preferences? Since we have access to the whole history of the blog site, we consecutively study the influentials in multiple 30-day windows and observe the temporal patterns of the influential bloggers.
- Are all the proposed parameters necessary?
- How can we extend the preliminary model? Are there any other parameters that can be incorporated?

Aggregating Similar Individual Blogs A fragmented Web entails the fragmented blogosphere. Given that a blogger has a social network, it seems sensible to start the search with the social network hoping that a similar blogger is the blogger's friend's friend (or so on). However, this seemingly simple idea is practically infeasible. It is a type of naïve link analysis that entails exhaustive search of the order $O(d^n)$, for n links and average degree d . Another reason that naïve link analysis cannot help much is that the Web is not a

Top 5 TUAW Bloggers	Top 5 Influential Bloggers
<i>Erica Sadun</i>	<i>Erica Sadun</i>
<i>Scott McNulty</i>	Dan Lurie
Mat Lu	<i>David Chartier</i>
<i>David Chartier</i>	<i>Scott McNulty</i>
Michael Rose	Laurie A. Duncan

Table 1: Two lists of the top 5 bloggers according to TUAW and our model, respectively.

random network. Its power law distribution suggests that more often than not, a blogger or group is in the Long Tail and not in the Short Head. In other words, they are largely disconnected as only those in the Short Head are well connected. Finding similar individual blogs on the blogosphere differs from classic data mining tasks. There are no typical training and test data. Hence, it requires innovative ways of evaluating and validating the end results.

5. RESULTS AND ANALYSIS

Here we first present the preliminary results for identifying the influential bloggers in a community blog. Then we present results for aggregating similar individual blogs so that these could be used as virtual communities and influential bloggers could be identified.

5.1 Identifying Influential Bloggers

We conduct our experiments on a real world community blog, The Unofficial Apple Weblog (TUAW - <http://www.tuaw.com>). Approximately, 15,000 posts are collected. We study the differences between active and influential bloggers, the evaluation of the model, temporal patterns of the influential bloggers, and parameter study.

Active and Influential Bloggers Many blog sites including TUAW publishes top 5 active bloggers for each month, based on their posting volume. We generate a list of top-5 bloggers using the preliminary model proposed in Section 2.1 and compare with the top 5 bloggers published at TUAW. We set the default values of all the weights as 1 assuming they are equally important. An in-depth study of these weights is presented in [3]. Table 1 presents the two lists of top 5 bloggers: the first column contains the top 5 bloggers published by TUAW and the second column lists the top 5 influential bloggers. Names in *italics* are the bloggers present in both lists. Three out of 5 TUAW top bloggers are also among the top 5 influential bloggers identified by our model. This set of bloggers suggests that some of the bloggers can be both active and influential. Some active bloggers are not influential and some influential bloggers are not active. By observation, there could be four types of bloggers: both active and influential (e.g., ‘Erica Sadun’, ‘David Chartier’, and ‘Scott McNulty’), active but non-influential (e.g., ‘Mat Lu’ and ‘Michael Rose’), influential but inactive (e.g., ‘Dan Lurie’ and ‘Laurie Duncan’), inactive and non-influential. Detailed statistics are presented in [3]

Evaluating the Model As we know, there is no training and testing data for us to evaluate the efficacy of the proposed model we look for an alternative to the ground truth. We resort to another Web2.0 site Digg (<http://www.digg.com/>) to provide a reference point. As people read articles or blog posts, they can give their votes in the form of digg and these votes are recorded on Digg servers. The higher the digg score for a blog post is, the more it is liked. Given the nature of

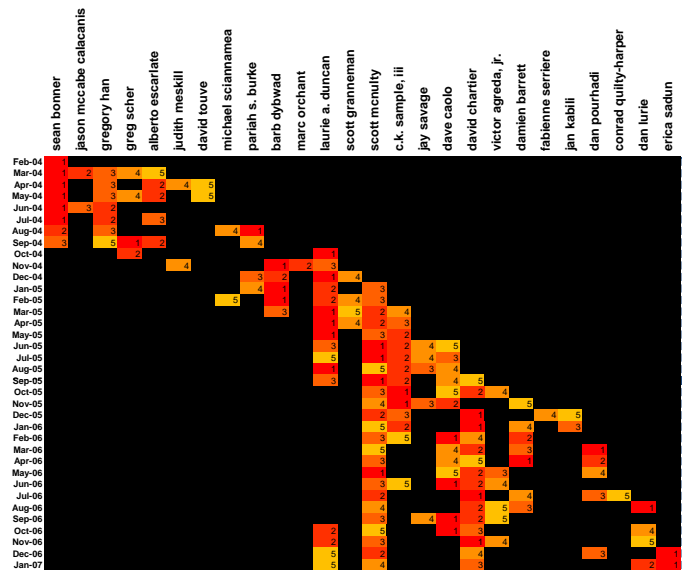


Figure 1: Influential Bloggers’ blogging behavior over the whole TUAW blog history.

Digg, a not-liked blog post will not be submitted thus will not appear in Digg. We take the four categories of bloggers, viz. 1. Active and Influential, 2. Inactive and Influential, 3. Active and Non-influential, and 4. Inactive and Non-influential and categorize their posts into S1, S2, S3, and S4, respectively. We rank the blog posts of each category based on the influence score and pick top 20 blog posts from each of the first three categories. We randomly pick 20 blog posts from the last category in which bloggers are neither active nor influential. Next we compare these four sets of 20 blog posts with the blog posts at Digg to study the overlap. The results are shown in Table 2. From the table, we can see that S1 has 17 out of 20 in the Digg set, and S4 has 0 or 1 found in the Digg set depending on randomization. The results show the differences among the four categories of bloggers and our model identifies the influentials whose blog posts are more liked than others according to Digg. For reference purposes, we also provide the distributions of Digg and TUAW blog posts in Tables 3 and 4, respectively. We observe from Tables 2, 3 and 4 that influential bloggers are more likely to be liked than active bloggers. More detailed discussion is included in [3].

Temporal Patterns of the Influential Bloggers For a blog site that has a reasonably long history, we can also study the temporal patterns of its influential bloggers. The blog site TUAW provides blogging data since February 2004. To study the temporal patterns of the influential bloggers we apply our model with a moving 30-day window with no overlap between two consecutive windows. In total, there are 26 influential bloggers during Feb.2004-Jan.2007. The temporal patterns of the influentials can be observed from a matrix in Figure 1. Influential bloggers are ordered according to the time they were recognized as influential vertically (column-wise), and the rows represent the progression of time. The (i, j) -th cell in this matrix stores the rank of the j th blogger in the i th time window. Black cells represent that the particular blogger was not among the top 5 for that time period. The color gradient represents rank of a influential blogger, a darker color representing a better rank.

Bloggers	Active	Inactive	Bloggers	Active	Inactive	Bloggers	Active	Inactive
Influential	S1: 17	S2: 7	Influential	S1: 71	S2: 14	Influential	S1: 327	S2: 42
Non-influential	S3: 3	S4: 0/1	Non-influential	S3: 8	S4: 7	Non-influential	S3: 131	S4: 35

Table 2: Intersection of Digg and top 20 from our model.

Table 3: Distribution of Digg blog posts.

Table 4: Distribution of TUAW blog posts.

- *Long-term influentials* They steadily maintain the status of being influential for a very long time, e.g. *Scott McNulty*. They can be considered “authority” in the community.
- *Average-term influentials* They maintain their influence status for 4-5 months, e.g. “Sean Bonner”, “Gregory Han”, and “Barb Dybward”.
- *Transient influentials* They are influential for a **very** short time period, e.g. *Michael Sciannamea*, *Fabienne Serriere* and *Dan Pourhadi*.
- *Burgeoning influentials* They are emerging as influential bloggers recently, e.g. *Dan Lurie* and *Erica Sadun*.

We conducted experiments to study the parameter relevance and pairwise correlation analysis to study the redundancy. According to the results none of the four proposed parameters was found redundant and the relevance of these parameters was found to obey the following order: inlinks > comments > outlinks > blog post length. We also study other parameters like rate of comments to gauge the influence of a blog post which is not a good feature to consider. More details on these experiments could be found in [3].

5.2 Aggregating Similar Individual Blogs

To evaluate the proposed model to aggregate similar individual blogs we conduct experiments on BlogCatalog. We collect individual blogs including the post tags, site tags, and category labels. The category labels have a hierarchical structure. We consider three variations of the data: top level (the labels of all the blog sites are abstracted to their top most parent level labels), all category (full hierarchical structure of the labels is considered), one node split. According to the distribution of blog sites in various top level labels, **Personal** has the largest number of blog sites. Detailed study is presented in [1]. Hence, we split **Personal** into its child labels, to reduce the skewed distribution of blog sites. Best results are obtained for all category dataset. More details could be found in [2]. We experiment with different link strength values (more details in [2]) and the best results were obtained for the value 5. We compared the WisClus based blog clustering approach (illustrated in Figure 2) with a baseline SVD based blog clustering using post content, blog post tags, and blog site tags. Due to space constraints the clustering results are available online at <http://www.public.asu.edu/~nagarwa6/baseline.jpg>. Upon observation¹:

1. Clusters obtained from baseline approach are too fragmented (lot of 2-member clusters) as compared to WisClus.
2. As a result, clusters are too focussed. This affects the insertions of new blog site later on. Cluster configurations are highly unstable in such a focussed clustering.
3. Several clusters from baseline clustering, have members whose blog site labels are semantically unrelated. For example, **bluemonkey jammies = Humor:Personal** and **emperoranton = SEO: Marketing** are clustered together. This is due to the susceptibility of vector space clustering to text noise, pre-

¹Pajek was used to create the visualizations.

dominantly found in blogs. Moreover, blogs are dynamic in nature with the blogger occasionally posting about different topics. However WisClus gives high-quality, semantically coherent clusters.

4. Several clusters obtained from baseline approach have members that have exactly the same labels. For example, the cluster with bloggers **emom** and **geraelindsey** have the same labels, i.e., **Small Business** and **Moms**. Clustering blog sites that have different yet related theme/topics are more helpful. WisClus generates clusters of blog sites with labels like, **Technology, Computers, Internet, and Technology>Gadgets**.

6. RELATED WORK

Here we present the state of existing solutions and discuss how the proposed solution is different or better as compared to existing approaches.

Ranking Blogs vs. Webpage Ranking The problem of ranking blog sites or bloggers differs from that of finding authoritative webpages. As pointed out in [12], *blog sites in the blogosphere are very sparsely linked and it is not suitable to rank blog sites using Web ranking algorithms* like PageRank [17] and HITS [11]. The Random Surfer model of webpage ranking algorithms [17] does not work well for sparsely linked structures. The temporal aspect is most significant in blog domain. While a webpage may acquire authority over time (its adjacency matrix gets denser), a blog post or a blogger’s influence diminishes over time. This is due to the fact that the adjacency matrix of blogs (considered as a graph) will get sparser as thousands of new sparsely-linked blog posts appear every day.

Influential Blog Sites How some blog sites influence the external world and within the blogosphere is studied by finding *influential blog sites*. It is orthogonal to the problem of identifying influential bloggers. Given the nature of the blogosphere, influential blog sites are few. A large number of non-influential sites belong to the long tail [4] where abundant new business opportunities can be explored. Our work is about identifying influential bloggers at a blog site regardless of the site being influential or not. We briefly review some work on influential blog sites. An interesting problem related to viral marketing [10] is how to maximize the total influence in the network (of blog sites) by selecting a fixed number of nodes in the network. A greedy approach can be adopted to select the most influential node in each iteration after removing the selected nodes. This greedy approach [8] outperforms PageRank, HITS and ranking by number of citations, and is robust in filtering splogs (spam blogs). Gruhl et al [7] study information diffusion of various topics in the blogosphere, drawing on the theory of infectious diseases via a general cascade model.

Community Identification There has been considerable amount of work on community construction based on link analysis in Web pages [11],[13]. However, the blogosphere has very sparse link structure [12] so it demands novel ways to synthesize communities from individual blogs.

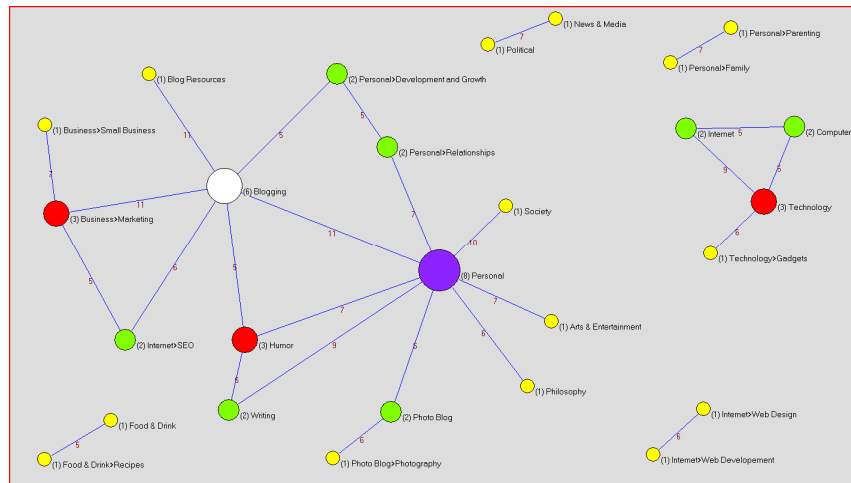


Figure 2: WisClus results for link strength ≥ 5 .

There have been some works that use content analysis of the blog posts to identify communities in the blogosphere [14]. Chin and Chignell [5] proposed a model for finding communities taking the blogging behavior of bloggers into account. However, these approaches suffer from the common disadvantages of text analysis i.e. high dimensionality and sparsity. Our proposed approach leverages the collective wisdom by means of category labels.

7. LOOKING AHEAD

Blogosphere is one of the fastest growing, social media. The virtual communities in the blogosphere are not constrained by physical proximity and allow for a new form of efficient communications. To better understand interesting activities happening in a virtual world, we need to study the blog communities and the influential bloggers in these communities. Nevertheless, the blogosphere consists of more individual blogs rather than community blogs. In this work, we present preliminary models to aggregate similar individual blogs to synthesize virtual communities and identify influential bloggers in blog communities.

Preliminary results for both the models look promising. However, more remains to be explored like, extending the influential blogger model to study parameters like readership, quality of comments (qualitative statistics). We plan to study trust aspect of the bloggers and add another dimension to the influential bloggers and make the system more robust and reliable by eliminating noise and splog issues.

8. ACKNOWLEDGEMENTS

This work is partially sponsored by AFOSR and ONR grants to Huan Liu.

9. REFERENCES

- [1] Nitin Agarwal et al. Searching for Familiar Strangers on Blogosphere: Problems and Challenges. In *NGDM*, 2007.
- [2] Nitin Agarwal et al. Clustering blogs with collective wisdom. In *ICWE*, 2008.
- [3] Nitin Agarwal et al. Identifying influential bloggers in a community. In *WSDM*, 2008.
- [4] Chris Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, 2006.
- [5] Alvin Chin and Mark Chignell. A social hypertext model for finding community in blogs. In *HYPERTEXT*, 2006.
- [6] Dan Gillmor. *We the Media: Grassroots Journalism by the People, for the People*. O'Reilly, 2006.
- [7] D. Gruhl et al. Information diffusion through blogspace. *SIGKDD Explor. Newsl.*, 6(2):43–52, 2004.
- [8] Akshay Java et al. Modeling the spread of influence on the blogosphere. In *WWW*, 2006.
- [9] Ed Keller and Jon Berry. *One American in ten tells the other nine how to vote, where to eat and, what to buy. They are The Influentials*. The Free Press, 2003.
- [10] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.
- [11] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- [12] Apostolos Kritikopoulos, Martha Sideri, and Iraklis Varlamis. Blogrank: ranking weblogs based on connectivity and similarity features. In *AAA-IDEA*, 2006.
- [13] Ravi Kumar et al. Trawling the web for emerging cyber communities. In *WWW*, 1999.
- [14] Ravi Kumar et al. On the bursty evolution of blogspace. In *WWW*, 2003.
- [15] Yu-Ru Lin et al. Splog detection using self-similarity analysis on blog temporal dynamics. In *AIRWeb*, 2007.
- [16] Tim O'Reilly. What is Web 2.0 - design patterns and business models for the next generation of software, 2005.
- [17] Lawrence Page et al. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.