Connecting Sparsely Distributed Similar Bloggers

Nitin Agarwal*, Huan Liu†, Shankara Subramanya†, John J. Salerno§ and Philip S. Yu¶
*University of Arkansas at Little Rock, Little Rock, AR 72204
Email: nxagarwal@ualr.edu
†Arizona State University, Tempe, AZ 85287
Email: {Huan.Liu,Shankara.Subramanya}@asu.edu
§Air Force Research Lab/IFEA, Rome, NY 13441
Email: John.Salerno@rl.af.mil
¶University of Illinois at Chicago, Chicago, IL 60607
Email: psyu@cs.uic.edu

***Abstract*—The nature of the Blogosphere determines that the majority of bloggers are only connected with a small number of fellow bloggers, and similar bloggers can be largely disconnected from each other. Aggregating them allows for cost-effective personalized services, targeted marketing, and exploration of new business opportunities. As most bloggers have only a small number of adjacent bloggers, the problem of aggregating similar bloggers presents challenges that demand novel algorithms of connecting the non-adjacent due to the fragmented distributions of bloggers. In this work, we define the problem, delineate its challenges, and present an approach that uses innovative ways to employ contextual information and collective wisdom to aggregate similar bloggers. A real-world blog directory is used for experiments. We demonstrate the efficacy of our approach, report findings, and discuss related issues and future work.**

I. Introduction

The unprecedented number of users participating in Web 2.0 [1] activities has generated enormous amounts of collective wisdom or open-source intelligence. Blogs are invigorating this process by encouraging the bloggers to document their ideas, thoughts, opinions, and views reverse-chronologically via so-called blog posts. As the blogosphere is part of the Web, it obeys the power law distribution. That is, the majority of bloggers occur in the long tail [2] - only a few bloggers are highly connected and a large number of bloggers are only connected with a relatively small number of bloggers.

Given a blogger b, we aim to find b's similar bloggers, and together, they can form a critical mass such that (1) the understanding of one blogger gives us a sensible and representative glimpse to others, (2) more data about similar bloggers can be collected for better customization (e.g., personalization and recommendation), (3) the nuances among them suggest new business opportunities, and (4) knowledge about them can facilitate predictive modeling and trend analysis in new product/market development. Connecting them to form a critical mass can not only potentially expand a blogger's social network (for job searching, special interest group formation, etc.), but also increase their Web prominence in targeted business campaigns. Aggregating similar bloggers can encourage participation due to the crowd effect [3]. Reputation and expression are among major motivations for bloggers to engage in activities; and shared interests will encourage them to participate more actively. People usually trust those with similar interests. Knowledge transfer or information flow among friends and acquaintances becomes smoother and more receptive.

Many bloggers in the long tail may share similar interests though they are largely disconnected. With Web 2.0, pervasive use of the Web technologies extends the long tail even longer, which makes personalization more important as well as challenging, thus demanding the use of emerging social media and the development of new algorithms. Aggregating such largely disconnected yet similar bloggers may lead to interesting applications such as identifying "Familiar Strangers" as mentioned in [4], [5].

The rest of the paper is organized as follows: Section II defines the problem and challenges. Section III motivates and describes our proposed approach. Section IV presents the details of data collection, experiments, and results. Section V discusses the related work with conclusions in Section VI.

II. Problems and Challenges

Most bloggers are associated with only a small group which entails a difficult task - finding other bloggers with similar interests. We delineate the problem of searching similar bloggers and present challenges.

A. Problem Statement

Given a blogger b, similar bloggers of b are a set of bloggers $B = \{b_1, b_2, ..., b_n\}$, who share common patterns as b, like blogging on similar topics [1]. Basically, every pair $\{b, b_j\}$ of bloggers, where $1 \leq j \leq n$, blog on similar topics or sharing distinct commonalities. $\{b, b_j\}$ are also non-adjacent - e.g., there are no connecting links in their blog posts or each one's presence in the other's social network. The social network information of b tells us the

friends of b who are directly connected to b. For $\{b, b_j\}$ to be *totally non-adjacent*, two conditions should hold true:

- b should not appear in b_j's social network, and
- b_j should not appear in b's social network.

Failing one of the two conditions would make them *partially non-adjacent*. For example, many adults in the US know of President Obama, but not vice versa. Henceforth, non-adjacent bloggers are *totally non-adjacent bloggers*.

The problem of finding similar bloggers can be formulated as: given a blogger b, identify 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 mentioned above. Similarity can be defined in terms of collectable statistics at blog sites (to be delineated in Section III-D.)

B. Challenges

Finding similar bloggers is essentially a problem of searching the long tail. Given a blogger's social network, search can be started from his(her) social network, assuming a similar blogger of b can be found in the social network of blogger c who is in b's social network. However, this seemingly simple idea is practically infeasible. It is typically a naïve link analysis that entails exhaustive search. Assuming each blogger has a social network of d friends, the search cost is $O(d^h)$ after exhausting bloggers who are h links away from the first blogger. It might very likely find similar bloggers, but incur the unbearable search cost. Another reason that naïve link analysis cannot help much is that the Web is not a random network. Its power law distribution suggests that a blogger is often in the long tail. In other words, they are largely disconnected as only those in the short head are well connected. Finding similar bloggers on Blogosphere differs from classic data mining tasks. There is no typically labeled dataset. Hence, it requires innovative ways of evaluating and validating the results. We will address these challenges to demonstrate the efficacy of various approaches and comparative findings.

III. Finding Similar Bloggers

In order to facilitate empirical study and performance evaluation, we use a blog site directory available at BlogCatalog as a microcosm for the blogosphere. There are two advantages of doing so: (1) We can use it as a concrete example to show generic features about blogosphere and explain what exactly this project aims to achieve; and (2) We can realistically evaluate the proposed approach in a semi-controlled environment. We first introduce BlogCatalog and then present our approach. Note that our approach is generic and can be adapted to any other blog structures.

A. BlogCatalog: A Blog Directory

Bloggers submit their blogs to BlogCatalog and specify the metadata such as, categories the blog is listed under, blog level tags, snippets of 5 most recent blog posts, and blog post level tags, for improved access to their blogs. A glimpse of BlogCatalogs directory structure is illustrated in Figure 1. The number of categories keeps increasing as new blog sites are submitted to BlogCatalog, although in a controlled fashion. At the time of writing, BlogCatalog had in total 56 categories. These categories could be denoted as $\{C_1, C_2, ..., C_n\}$ and $n = 56$. Blog sites could be denoted as $\{s_1, s_2, ..., s_m\}$. Each site is authored by a blogger denoted by $\{b_1, b_2, ..., b_m\}$. Bloggers' friends are collected using BlogCatalog's API.

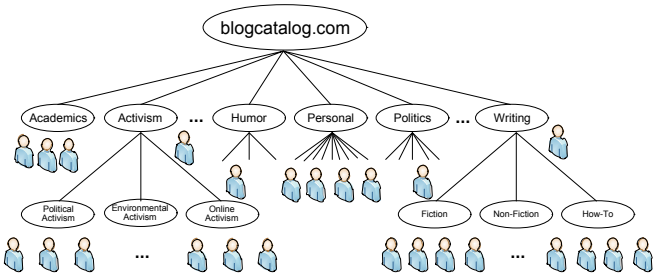

Figure 1. BlogCatalog's Directory Structure.

B. A Collective Wisdom based Search Approach

One straightforward way is to use the social network information to connect bloggers via links. As mentioned above, one can traverse friend's link, friend's-friend's links, and so on to find and connect similar bloggers. It is impractical to perform such an exhaustive search of similar bloggers because of the fragmented blogosphere. Moreover, there are simply too many bloggers to search and there are a relatively small number of bloggers out there who share similar interest for a blogger. This naïve link-based approach needs an alternative.

Since the blogosphere normally organizes the blog sites under categories, as specified by the blogger themselves, one intuitive solution is to search similar bloggers of b in the categories b is associated with. This approach could be termed as the *naïve taxonomy based search approach*. However, these categories may not be suitable for our purpose: as illustrated in Figure 10 later, bloggers often use "Personal" as the category descriptor for varied interests. Using it may not be helpful in capturing the nuances of bloggers' interests and we need to refine the category descriptor by identifying and aggregating the related categories. This is referred as the topic irregularity problem where bloggers use the same category descriptor to define their blog site which in fact contains blog posts of varied interests. Connecting similar categories would need to cater to these nuanced interests which otherwise are not expressed by the category descriptors. That would require that different categories with similar themes be connected even when a blogger does not list his/her blog sites under all these categories. For example,

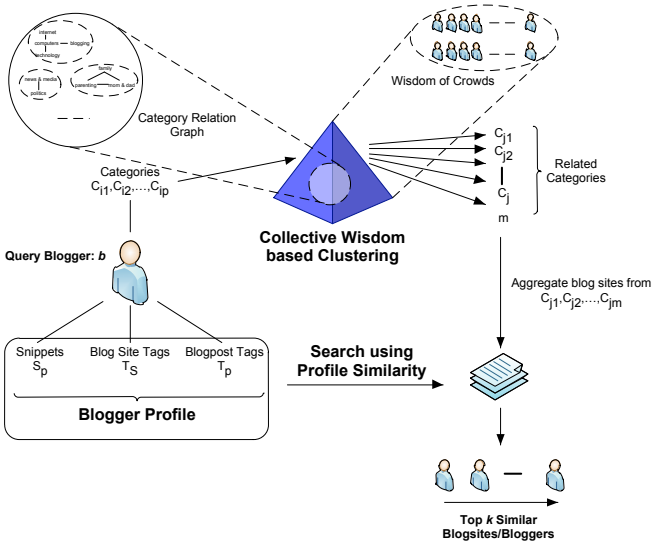

Figure 2. A diagrammatic sketch of Collective Wisdom based Search.

a blog site named "Words From Iraq" is listed under "Iraq and Society"; another blog site, "Iraq's Inconvenient Truth", is listed under "Political", and "News & Media". Although they are related blog sites, they are listed under completely different categories. Connecting these categories through some sort of link or relation would augment the search space for relevant results. Now the issue is what categories are most likely connected. Since each individual can behave very differently, the information of an individual blogger cannot help find a solution. Given the nature of blogging, however, some new information emerges from collective blogging, or collective wisdom.

Our proposed approach takes advantage of the collective wisdom in search of similar bloggers, as illustrated in Figure 2 - given a query blogger b as input, outputs top-k similar bloggers. First we briefly list the essential steps in our approach and then explain each one in detail.

1) Find related categories to that of the query blogger b;
   a. For a query blogger b, identify the categories (s)he has used to describe his/her blog.
   b. Identify other related categories of blogger b for search. Collective wisdom can be leveraged to identify possible categories in which similar bloggers of b may be found (explained in Section III-C).
2) Find similar bloggers to the query blogger b from the bloggers in these categories;
   a. Construct profile for the query blogger b and bloggers from the categories identified in the previous step using their respective blog site level tags, blog post level tags, and snippets (explained in Section III-D1).
   b. Search for b's similar bloggers in the categories identified in previous step based on blogger profiles by transforming this high-dimensional feature vector into concept space (explained in Section III-D2).

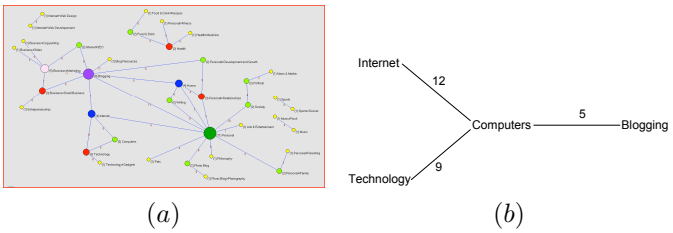

Figure 3. Identifying similar categories using collective wisdom. (a) Full view of the obtained CRG, (b) Snippet of CRG illustrating similarity relationship between 4 categories.

Our proposed search algorithm for similar bloggers is referred to as the *collective wisdom based search*, or CWS. We elaborate the algorithmic details in the next two subsections. Later in Section IV-C1, we compare CWS with the naïve taxonomy based search approach.

C. Category Similarity

A naïve solution to finding and connecting similar categories is to simply treat the categories as term vector created by aggregating the tags from the blog sites and blog posts of that category and then computing a pairwise cosine similarity metric. Mathematically, if category C_i is represented by the term vector v_i and category C_j is represented by the term vector v_j, then the similarity between C_i and C_j represented as $Sim(C_i, C_j)$ is defined as:

$$Sim(C_i, C_j) = \frac{v_i \times v_j}{|v_i||v_j|} \quad (1)$$

Here the term vectors v_i and v_j are constructed by aggregating the blog site level tags and blog post level tags for all the blogs listed under categories C_i and C_j, respectively. This document term representation of the categories based on the tags can be high-dimensional and extremely sparse. To find similar categories using this solution can be extremely inefficient given the massiveness of the data. So we explore alternative novel ways to identify similar categories.

Bloggers provide the categories a blog site belongs to. If a lot of bloggers use multiple categories simultaneously to describe the theme of their blog sites, then we can fairly assume that these categories could be related. This results in a category relation graph (CRG) as shown in Figure 3(a). Figure 3(b) depicts a small snippet of the full CRG. Here, categories 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 categories is termed as *Link Strength*, as depicted by the numbers shown on the edges of the category relation graph in Figure 3(b). Using this category relation graph, similar categories can be linked or connected. We experiment with different thresholds for the link strength in Section IV-B3. More experiments

comparing collective wisdom based approach and other techniques could not be presented here due to space limitations; however, interested readers could follow [6] for these results and more details.

The collective wisdom based approach is time sensitive resulting in dynamic CRG. Every time new blog site appears and blogger specifies different categories, CRG might change. We observe different CRG for different crawls of BlogCatalog (more in Section IV-C2), which verifies the dynamism of collective wisdom. Since Blogosphere emphasizes on the freshness of the content, similar dynamics is expected from collective wisdom based approach.

D. Blogger Similarity

Searching the categories identified through CRG to find similar bloggers requires a notion of similarity between bloggers. Estimating similarity between bloggers presents challenges like, what information could be used to represent bloggers efficiently; how to represent the bloggers in a way that avoids issues like high-dimensionality and sparsity; and how to quantify the notion of similarity. One way is to construct blogger profiles such that similarity between profiles can be compared. We first discuss blogger profile construction and then present similarity computation.

1) Blogger Profiles: A blogger's profile can be constructed using multiple information sources: blog post level tags, blog site level tags, and blog post snippets. Blog post level tags are highly dynamic in nature. We combine all the tags for all the blog posts for each blogger available at BlogCatalog and construct a matrix, T_p, where rows represent blog post level tags and columns represent bloggers. Blog site level tags are less dynamic as compared to blog post level tags but they give precise information about the blog site. The most dominant theme of the blog site can be gauged with these tags [7]. We construct a matrix, T_s, where rows represent blog site level tags and columns represent bloggers.

The third source of information is the blog post snippets. They are pieces of texts that require some preprocessing (stopword elimination and stemming [8]) before they can be used. All the words from all the snippets for each blogger are then combined and a matrix, S, is constructed where rows represent terms and columns represent bloggers.

Snippets are less sparse and noisier, which is treated differently from tags which are sparser and less noisy. To integrate the two types of tag data (post and site) we use a linear combination of the information sources. The tag matrices T_p and T_s are combined to get a single tag matrix T.

$$T = \beta \times T_s + (1 - \beta) \times T_p \tag{2}$$

Here β controls contribution of site-level tags (T_s) and post level tags (T_p) in the overall tag matrix, T. In the experiments, we consider $\beta = 0.5$, i.e., equal contribution of both T_s and T_p. However, β can be tuned to adapt to different datasets depending on what type of tags are more readily available and reliable. Bloggers' profiles could be collectively represented as a set depicted by $\{T, S\}$. A blogger x_j's profile is represented by the j-th column of these two matrices. Precisely, x_j's profile is represented as $\{T_j, S_j\}$.

2) Profile Similarity: The profile thus described can be used to compute pairwise similarity between a blogger, x_j and other bloggers in the identified categories through CRG. Cosine similarity function could be used to compute similarity between bloggers. However, the profiles are very sparse and high-dimensional. Moreover, naïve term matching does not include the context information and does not perform well for problems like polysemy. Hence we use Latent Semantic Analysis (LSA) [9] to evaluate similarity between bloggers.

Let X denote the blogger term matrix, where cell (i, j) denotes the frequency of term i for blogger j. Each column in X represents a blogger and the terms in rows represent his/her profile. For instance, x_j denotes the blogger j. A cosine similarity between the column vectors of X would give a similarity between the bloggers, but due to above mentioned challenges this approach is highly inefficient. Hence we perform LSA transformation of matrix X to a concept space which is less sparse and low-dimensional as,

$$X = U\Sigma V^T \tag{3}$$

where U denotes the transformed profile terms of the bloggers in concept space, Σ is a diagonal matrix of the singular vectors, and V^T denotes the transformed bloggers' profiles in concept space. A cosine similarity between the columns in V^T matrix would give us similar bloggers in concept space. If X is $d \times m$ matrix, where m denotes the total number of bloggers and d denotes the total number of terms in the profiles, then matrix U is $d \times \ell$, Σ is $\ell \times \ell$ and V^T is $\ell \times m$. Here ℓ is the total number of concepts in the reduced space. We can select the k largest singular values which would further reduce the concept space from ℓ to k. This gives the rank k approximation with smallest error.

The above described LSA transformation is applied to both the tag matrix (T) and the snippet matrix (S).

$$T = U_t \Sigma_t V_t^T, \quad S = U_s \Sigma_s V_s^T \tag{4}$$

The lower dimensional representation of V_t and V_s denoted by V_{tk} and V_{sk} respectively gives the two components of the blogger profiles in the reduced concept space for first k eigenvectors. The j-th row of V_{tk} and the j-th row of V_{sk} denoted by v_{tj} and v_{sj} forms the blogger profile for the blogger x_j in reduced concept space. That is, profile of blogger, $x_j = \{v_{tj}, v_{sj}\}$. Here v_{tj} represents the 'tag component' of the profile and v_{sj} represents the 'snippet component' of the profile. Based on bloggers' profiles, we compare their profiles to find the blogger similarity. We use

cosine similarity as the similarity measure for this comparison. So given two bloggers x_j and x_r, whose profiles are represented as $\{v_{tj}, v_{sj}\}$ and $\{v_{tr}, v_{sr}\}$, respectively, we compute their similarity:

- Tag Similarity (Sim_{tag}) is the cosine similarity between tag components of the two profiles.

$$Sim_{tag}(x_j, x_r) = \frac{v_{tj} \times v_{tr}}{|v_{tj}||v_{tr}|} \quad (5)$$

- Snippet Similarity ($Sim_{snippet}$) is the cosine similarity between snippet components of the two profiles.

$$Sim_{snippet}(x_j, x_r) = \frac{v_{sj} \times v_{sr}}{|v_{sj}||v_{sr}|} \quad (6)$$

We can then combine the two similarity values to get the final similarity between the two bloggers as:

$$Sim(x_j, x_r) = \alpha Sim_{tag}(x_j, x_r) + (1-\alpha) Sim_{snippet}(x_j, x_r) \quad (7)$$

Here α is a tunable parameter. Depending upon the kind of profile similarity needed and the type of information available, weights could be adjusted to emphasize more on dynamic information sources like snippets and blog post tags or more stable sources like blog site tags. We evaluate the performance for different combinations of tags and snippets in Section IV-B1.

The advantage of a linear model as shown above is its extendibility. We can later on add other information sources as they become available. For instance, we can add category level tags derived from blog sites and blog posts of that category. These blog sites and blog posts could be a random sample of the category or from influential bloggers [10], [11] of a particular category. Influential bloggers are used as the representatives of the category. Blog site category level tags tell us the broad topics under which the blog site can be categorized. These tags are more stable than blog post tags.

IV. Experiments and Results

CWS presents a palpable way of identifying similar long tail bloggers in Blogosphere. In this section we present experiments for our approach using collective wisdom to search similar bloggers and answer the following questions:

- Does CWS give better search results as compared to the naïve taxonomy based search in terms of accuracy?
- Does the CRG capture the dynamics of the collective wisdom as dataset changes?
- How does CWS compare with naïve taxonomy based approach in terms of search space reduction?

Before we present the experiments to study and answer the above-mentioned questions we briefly explain the experiment setup including data collection, validation strategy, and training and testing phases.

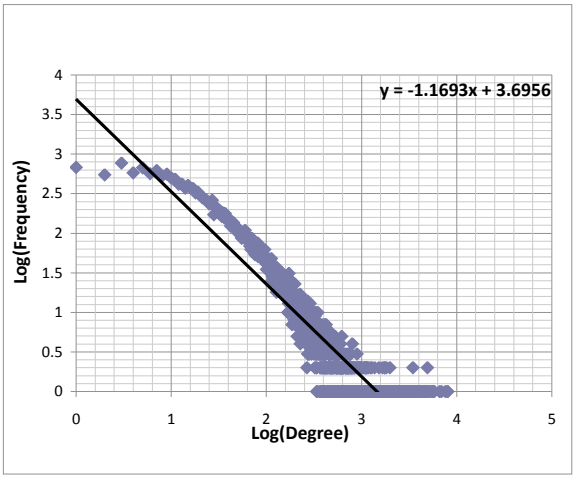

Figure 4. Log-log plot of the bloggers social network degree distribution.

A. Experiment Setup

1) Data Collection: BlogCatalog data was collected using 4 bloggers and their social network as starting points. These 4 bloggers belong to the most popular categories (i.e., having largest number of blog sites) at BlogCatalog. Using this procedure we crawled 12,308 bloggers. For each blogger thus crawled (uniquely identified by their blogger IDs), we collect their blog site level information like, blog site URL, blog site title, blog site categories, blog site tags, number of hits, number of views, blog site rank; blog post level information like, blog post tags, blog post snippets, date of posting; and the bloggers social network information, i.e., his/her friends. We plot the bloggers crawled vs. the size of their social network also known as their degree in Figure 4, which shows an expected power law distribution with exponent equal to 1.1693. It clearly shows that there are few bloggers that are highly connected and an overwhelming number of bloggers that are disconnected.

2) Validation Strategy: Since the ground truth about similar bloggers is not available, we need to construct one. Given a query blogger we search exhaustively in the collected dataset for similar bloggers by constructing blogger profiles and computing similarity using LSA as explained in Section III-D. We record the top 3 results (bloggers) returned by the exhaustive search and treat it as the best available ground truth. On the other hand CWS constraints the search space by searching only the related categories identified by CRG. We return top 3 results (bloggers) from CWS and report mean average precision (MAP), defined as:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk}) \quad (8)$$

Here Q refers to a set of queries (bloggers), $m_j = 3$ (top 3 results), and $P(R_{jk})$ refers to precision after each relevant

result (blogger) is added to the result set. $P(R_{jk})$ implies the precision obtained when the recall is equal to R_{jk}. MAP incorporates both precision and recall perspectives in the results and generates one value for the whole query set which is easier for comparison. We exploit collective wisdom and validate against the exhaustive search.

3) Experiment Setup: At any given time it is impossible to collect all the data from the blogosphere in real time. So in order to simulate this constraint we do not use the complete data for computing the similarity matrix. We divide the data into training and testing as shown in Table I. 80% of the collected data is treated as training data and the rest 20% is treated as the test data. Training data is used to construct the transformed concept space of the bloggers' profiles. Bloggers in testing data are used as query bloggers (Q) and transformed to concept space using the transformation developed through the training data.

Table I
DIVISION OF DATA INTO TRAINING AND TEST.

Total bloggers in the dataset	**Bloggers used as training data**	**Bloggers used as test data or query**
12,308	9,846	2,462

Let T and S denote the blogger term matrices for the tag and snippet sources respectively, obtained from bloggers in the training data set. These matrices are decomposed as mentioned previously in Section III-D2,

$$T = U_t \Sigma_t V_t^T, \quad S = U_s \Sigma_s V_s^T \tag{9}$$

Let U_{tk} , Σ_{tk} , V_{tk} , U_{sk} , Σ_{sk} , and V_{sk} represent the rank k approximation of the corresponding matrices. Let xtj and xsj represent the blogger term vectors for the blogger xj from tag and snippet sources respectively. Then we have,

$$x_{tj} = U_{tk} \Sigma_{tk} x'_{tj}, \quad x_{sj} = U_{sk} \Sigma_{sk} x'_{sj} \tag{10}$$

where x'_{tj} and x'_{sj} are the blogger profiles in the transformed concept space. The above equations can be modified as shown below to obtain the transformed concept space profile vector given the original blogger profile:

$$x'_{tj} = \Sigma_{tk}^{-1} U_{tk}^T x_{tj}, \quad x'_{sj} = \Sigma_{sk}^{-1} U_{sk}^T x_{sj} \tag{11}$$

Here, $(\Sigma_{tk}^{-1} U_{tk}^T)$ and $(\Sigma_{sk}^{-1} U_{sk}^T)$ are the transformation models from original space to concept space. The profile of the blogger using the transformed vectors is therefore given as,

$$x_j = \{x'_{tj}, x'_{sj}\} \tag{12}$$

For a query blogger q, its transformed concept space vectors for the tag and snippet sources, using the learned transformation model, can be computed as:

$$q'_t = (\Sigma_{tk}^{-1} U_{tk}^T) q_t, \quad q'_s = (\Sigma_{sk}^{-1} U_{sk}^T) q_s \tag{13}$$

Using these transformed tag and the snippet vectors we can construct the profile of the query blogger q as,

$$q = \{q'_t, q'_s\} \tag{14}$$

Once the reduced concept space profile of the bloggers is constructed the similarity between the blogger q (of the test set) and the blogger x_j (of the training set) is calculated. For the blogger q, top three similar bloggers from the training set are selected and returned as results.

B. Model Parameters

Our approach depends on parameters like, weights for combining similarity from blog site level tags and blog post level tags (α), and blog post snippets ($1 - \alpha$), number of eigenvectors, and threshold for category link strength.

1) Weights for Similarity: α: For experimenting with weights, we considered three configurations: (1) Tags only ($\alpha = 1$), (2) Snippets only ($\alpha = 0$), and (3) Both tag and snippets ($\alpha = 0.7$)[1]. MAP results for all the 3 configurations are shown in Figure 5. It is clear from the Figure 5 that using tags and snippets together achieves better MAP as compared to using only tags or snippets. Here the improvement in MAP by using tags and snippets together is not significant as compared to tags only. This is because the blog sites are well labeled. In the case where the blog sites are poorly labeled (either missing labels or noisy labels) using snippets would help. The proposed model provides flexibility to tune the contribution of tags and snippets through α which can be adjusted according to the dataset. For the rest of the experiments we pick the configuration 3, i.e., using both tags and snippets to build a blogger's profile.

2) Number of Eigenvectors: Next we experiment with different eigenvectors (k = 10, 25, 50, and 100) for LSA. We report MAP results for different values of eigenvectors with category link strength of 4 and higher in Figure 6. It is clear from Figure 6 that both 10 and 25 eigenvectors give the best MAP results (91.164%), however varying different threshold values for category link strength, 25 eigenvectors performs more robust than 10 eigenvectors. As the number of eigenvectors increases beyond 25, MAP decreases because the concept space starts getting noisy. So for rest of the experiments we chose the first 25 eigenvectors for LSA.

3) Threshold for Link Strength: The MAP result for our approach with different threshold values of category link strength is shown in Figure 7. The threshold for category link strength values vary from 3 to 15. We do not consider threshold values less than 3, because link strength less than 3 could be accidental. From this figure we observe that at threshold value of 4, the MAP stabilizes. At other values the MAP is either unstable or very low. So 4 is chosen as the link strength threshold for constructing the CRG.

[1]This value of α gives the best result. Due to space constraints we do not include the results for other values of α.

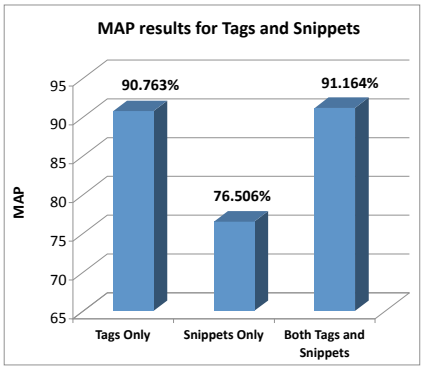

Figure 5. MAP for different α.

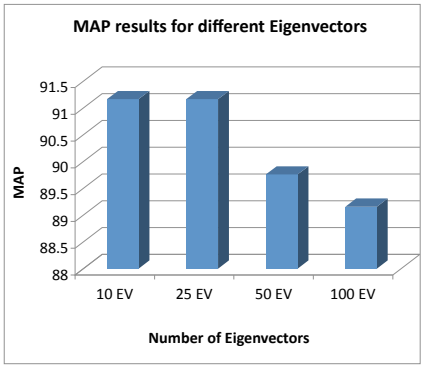

Figure 6. MAP for different eigenvectors.

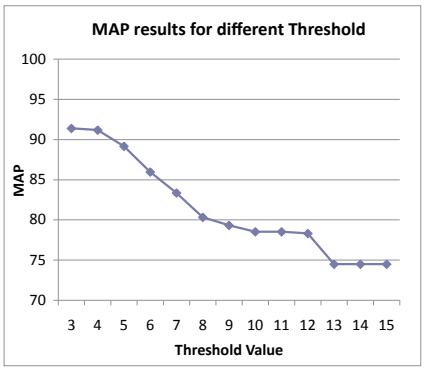

Figure 7. MAP for different link strength threshold.

C. Impact of Using Collective Wisdom

In this section we design experiments to evaluate the effect of using collective wisdom in search of similar bloggers. First, we compare CWS with the taxonomy-based search that does not employ collective wisdom. We compare the MAP values for both approaches. Second, we explore if CWS can capture the dynamics of category relations in terms of the category relation graph. Last, we study the search space reduction using CWS.

1) Performance of CWS: We perform experiments to compare CWS with the naïve taxonomy based search approach. CWS generates CRG which can be used to expand the search space for a given query blogger. Whereas, naïve taxonomy-based approach searches within the category of the query blogger. This approach does not use collective wisdom to identify similar categories. A comparison between these two approaches tells us whether collective wisdom is helpful or not. Results are shown in Figure 8. MAP is computed for both the approaches using the exhaustive search results as the ground truth. We report the results for different number of eigenvectors. It is clear from Figure 8 that CWS outperforms naïve taxonomy based search approach. Best value for MAP for naïve taxonomy based search approach is 77.254% at 10EV, CWS achieved MAP of 91.164% at 10EV and 25EV. This shows that expanding the search space by using CRG generated by collective wisdom improves the performance. Detailed search space analysis is presented in Section IV-C3.

2) Dynamics of Collective Wisdom: Since Blogosphere is a very dynamic environment, we expect similar characteristics to be exhibited by the collective wisdom that emerges from Blogosphere. Basically, the category relations can change and evolve over time. In order to study the dynamic characteristics of collective wisdom, we look at the changes resulted from data containing different number of bloggers i.e., 7023, 10642, 11947, and 12308 bloggers. Increase in data causes the change of collective wisdom which can be captured in category relation graphs (CRGs).

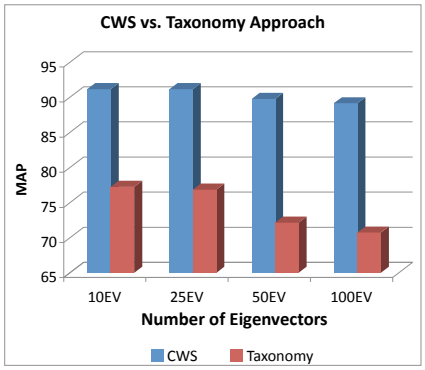

Figure 8. Comparison between CWS and Taxonomy based search approach.

This suggests how the tagging behaviors of various bloggers change over time. To study the dynamics of the collective wisdom we visualize the CRGs obtained from the data containing 10,462 bloggers and 12,308 bloggers, depicted in Figure 9 as 10k and 12k, respectively.

Comparing the CRGs of data with 10k and 12k bloggers in Figure 9, we observe that "Computers - Science - Technology" and "Internet - Business -Blogging - Blog Resource" are merged in 12k. This is due to a new link that emerges between "Technology" and "Internet". Similarly we notice a new link between "Society" and "News & Media" categories. "Political - News & Media - Society" in SB2 expands to "Political - News & Media - Society - Humor - Writing" in 12k. This transformation connects categories, "Society" with "Humor" and "Humor" with "Writing". Another instance of expansion is "Coaching - Education & Training" in SB2 to "Career & Jobs - Coaching - Education & Training" in SB4. This creates a new link between "Career & Jobs" and "Coaching". We also notice changes to existing links like "Travel - Vacation" which transforms to "Photo Blog - Travel - Vacation". This establishes a new link between "Photo Blog" and "Travel" categories. Some

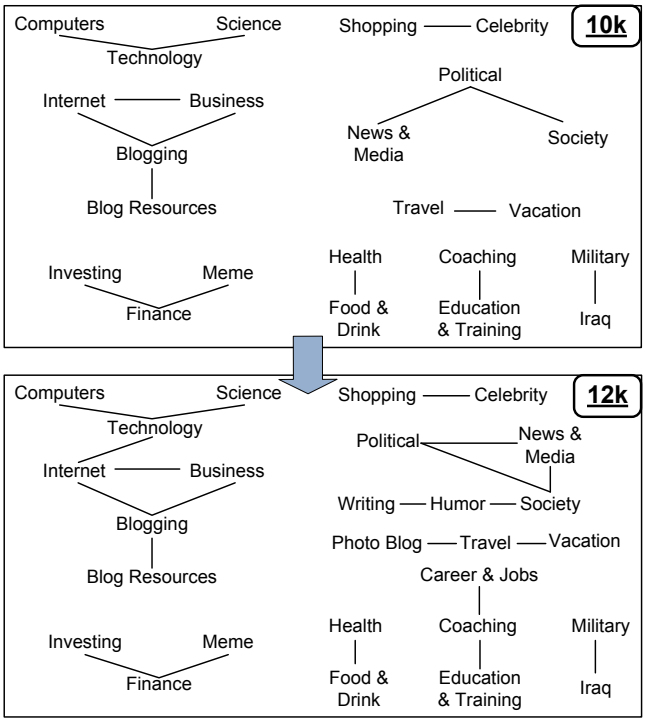

Figure 9. CRGs for different datasets containing 10,642 bloggers (10k) and 12,308 bloggers (12k).

of the category links remain unchanged in 10k and 12k like, "Investing - Finance - Meme", "Shopping - Celebrity", "Health - Food & Drink", "Military - Iraq". Some category relations like "Crafts - Arts & Artists", "Places - History - Academics", "Philosophy - Religion" remain intact as data increases.

3) Search Space Analysis: Here we compare the search space complexity of CWS and naïve taxonomy based approach to that of the exhaustive search method. Search space reduction is computed as:

$$\Gamma_j = \frac{\sum_k B(C_k^k)}{\sum_{i=1}^{n} B(C_i)}, \quad SSS = 1 - \frac{\sum_{j=1}^{|Q|} \Gamma_j}{|Q|} \tag{15}$$

Here Γ_j represents the ratio of search space complexity of the proposed approach (Naïve taxonomy based approach or CWS) to the exhaustive search for query blogger j. C_{kj} denotes the candidate categories for query blogger j to be searched. For naïve taxonomy based approach the candidate categories are simply the categories of the query blogger, whereas, for CWS the candidate categories include the query blogger j's categories as well the related categories identified by the CRG. $B(C_i)$ denotes the total number of blog sites in category C_i. n is the total number of categories in the dataset. Search space reduction (SSR) denotes the average reduction in search space for each blogger in the query set Q. $|Q|$ denotes the size of the query blogger set.

We obtain 91.164% MAP after 28.723% search space reduction, i.e., we search approximately 71.276% of the data on average for all the query bloggers. This does not seem quite significant; however, analyzing the data reveals interesting findings:

1. The taxonomy structure is not used by the bloggers to its full potential. This is demonstrated by the following observation: if we simply search the categories under which the query blogger has listed his blog (as in naïve taxonomy based approach), we still have to search 57.916% of the data on average for all the query bloggers. Increasing the search space by a margin of 13.36% from naïve taxonomy based search approach to CWS by including related categories the search performance increased to 91.164% from 77.254%.

2. If the BlogCatalog taxonomy structure is used efficiently then the search space can be further reduced, as evident by the following observation: the BlogCatalog dataset has highly unbalanced distribution with more than 33% of bloggers in Personal category and more than 20% of bloggers in Blogging category, as evident from Figure 10. This is because these categories are very generic[2]. Though they have more specific subcategories but bloggers tend to submit their blogs under these generic categories either they are ignorant of the subcategory structure (also because the taxonomy structure is highly dynamic and keeps evolving), or because they are lazy to submit their blogs to more focused or refined categories or subcategories, or bloggers keep using the same category labels to the new blog posts or new blog sites as the old ones. This problem is also referred as path dependence [12]. If we assume that both these most popular categories are further divided into 3 subcategories and bloggers are well versed with this structure then we can save at least 22% (=33*2/3) of search space from Personal and at least 12% (=20*2/3) from Blogging. This accounts for an additional search space reduction of at least 34%. So the total search space reduction on adding the search space reduction from CWS would be at least 28.723% + 34% = 62.723%. This is a conservative estimate since we assume only 3 subcategories under the two most popular categories that bloggers use, i.e. Personal and Blogging. However in reality, Personal has 12 subcategories and Blogging has 5 subcategories. To justify the analysis, we performed a controlled dataset experiment. Here we collected the dataset that did not contain blogs that had either Personal or Blogging listed as one of the categories. This ensures the distribution is not highly unbalanced. We obtained 51.526% search space reduction using CWS, which is in accordance to the conservative estimate presented above in the analysis.

Although bloggers do not use the taxonomy structure efficiently due to the reasons mentioned above, it is possible to

[2]For the sake of space constraint and the analysis presented here, we limit the categories in this chart that have at least 1000 blog sites.

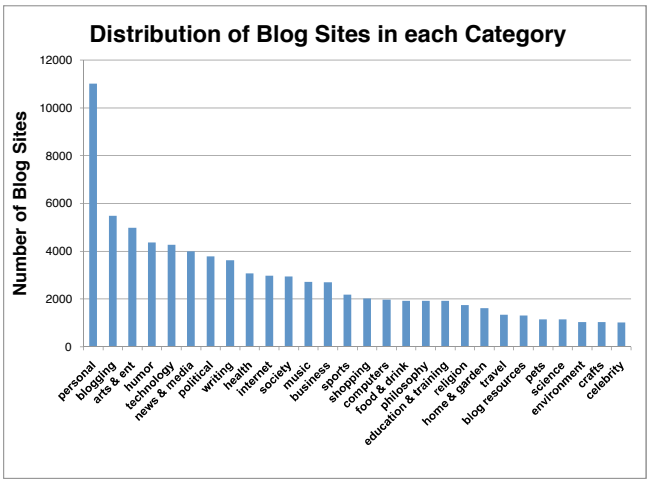

Figure 10. Number of Blog Sites in different categories in BlogCatalog.

Table II
MAP RESULTS FOR 10642, 12308, AND 12745 BLOGGERS.

Number of Bloggers	10,642	12,308	12,745
MAP	74.297%	91.164%	93.373%

provide reasonable solution to this problem. We can treat this as a multiclass classification problem. Subcategories could be treated as the class labels and blogs already submitted to the subcategories could be treated as the training set. Now the blogs that are submitted to one of the most frequently used categories (C_k) could be treated as test instances and classified into one of the subcategories of C_k. However, this is out of the scope of this paper and could be pursued as a future research direction.

Next we perform an experiment to study the effect of increasing the number of starting bloggers to crawl the data. We compared the MAP of CWS for 10,642, 12,308, and 12,745 blogger datasets obtained by crawling 2, 4, and 6 bloggers as starting points. Table II shows the MAP results for these datasets. It is clear from the results that dataset with 12,745 bloggers performs better than 10,642 and 12,308 blogger datasets. However, the performance gain from 10,642 blogger dataset to 12,308 blogger dataset is much steeper than from 12,308 blogger dataset to 12,745 blogger dataset. This shows that collecting more data certainly makes the similarity matrix for LSA more robust and accurate; however, beyond a certain point collecting more data does not help too much.

V. RELATED WORK

One line of related research is mining relationships between people based on the text they have written using LDA and its variations [13], [14]. These approaches develop topic models on the documents submitted by the authors. Authors may produce several documents often with coauthors, which makes it unclear how the topics generated for these documents might be used to describe the interests of the authors. Moreover, it is also a challenge to learn the parameters in these approaches even though well-established approximation techniques exist. Considering the limitations of author-topic model based approaches to identify latent relations, authors in [15] train an SVM classifier to predict the topics for bloggers from external topic taxonomies. Based on the topic similarity and further refined by the cosine similarity of actual blog content, similar bloggers can be recommended. As topic taxonomies keep evolving, it requires retraining the classifier that adds to the complexity of the solution. Moreover, detecting bloggers' true interests in some of their writings is another challenge. We instead rely on the bloggers' annotations to identify the set of topics or categories bloggers are interested - a type of collective wisdom generated by fellow bloggers, including related categories in the search for similar bloggers.

Some typical problems in Social Network Analysis (SNA) include discovering groups of individuals sharing the same properties using the connectivity properties of networks [16]. Since this does not consider the textual information of the entities, it limits the applications of SNA. Moreover, the sparseness of links among blogs can greatly limit these approaches [17]. The typical keyword search also encounters the power law distribution when searching for documents. For a given query, the search algorithm like PageRank [18] finds a large number of matching results. To handle the large number of results, the algorithm ranks the documents according to their authoritativeness, with highly connected documents ranked at the top. The problem is that those sparsely distributed bloggers we aim to find are often buried or hidden in the middle or the end of the extensive returned results, a.k.a. in the long tail.

Researchers have also focused on the problem of blog search and mining relevant content from the random and chaotic information content in Blogosphere. Due to the infancy of the blog search engines [19] more complex models and information retrieval techniques are needed to increase the quality and accuracy of the search. Authors in [20] use PLSA to segment the blogs into various topics and identify high usage keywords in these topics. Each blog is converted to concept space using LSI and cosine similarity between query keywords and blogs is used to compute the search results. However, this differs from the work proposed in the paper in two aspects: (1) Authors in [20] focus on business and corporate blogs which are less extraneous and contains more useful information than personal blogs, which we focus in our work; and (2) the work proposed in this paper searches for similar bloggers given a blogger which is different than a blog search engine which works on a keyword-query based model. Moreover, we use the CRG obtained from collective wisdom to segment the blogs. Identifying similar bloggers

differs from typical blog clustering.

Identifying similar bloggers could be remotely compared with blog clustering. Blog clustering [21] attempts to organize blogs or bloggers into meaningful or semantically related groups. Authors in [22]; however; use blog clustering to improve the annotation of blogs by automatically suggesting tags. Our work focuses on finding bloggers that are most similar to a given blogger. In other words, our approach tries to find k nearest neighbors of a given blogger instead of finding k clusters of bloggers.

VI. Conclusions

The sparsely linked blogosphere presents a new problem - searching for similar bloggers in the long tail of Blogosphere, in order to discover and connect niches in the long tail where the majority of bloggers are largely disconnected. This problem raises many technical challenges. We formally define the problem, provide working definitions, illustrate the challenges, and enumerate some potential solutions and baseline approaches. We then propose and develop a novel solution based on local information and global context that emerges from collective wisdom of bloggers. To facilitate the empirical study, we attempt to find a representative microcosm by sampling a real-world blog directory to approximate the blogosphere, retaining generic features. Further work will definitely lead to many innovative approaches for searching and connecting similar bloggers in the long tail.

Acknowledgment

This work is in part supported by AFOSR grant FA95500810132, ONR grants N000140810477, N000140910165, and the U.S. National Science Foundation grant IIS-0905215. We would like to thank Magdiel Galan for helping us with illustrations.

References

[1] S. Nigel and B. Tim, "Web science: Studying the internet to protect our future," Scientific American Magazine, September 2008.

[2] C. Anderson and M. Andersson, *The long tail: Why the future of business is selling less of more*. Hyperion New York, 2006.

[3] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "Structure and evolution of blogspace," *Comm. of the ACM*, vol. 47, no. 12, pp. 35–39, 2004.

[4] N. Agarwal, H. Liu, J. J. Salerno, and P. S. Yu, "Searching for Familiar Strangers on Blogosphere: Problems and Challenges," in *NSF Symposium on Next-Generation Data Mining and Cyber-enabled Discovery and Innovation (NGDM)*, 2007.

[5] N. Agarwal, H. Liu, S. Murthy, A. Sen, and X. Wang, "A social identity approach to identify familiar strangers in a social network," in *Proceedings of the 3rd International AAAI Conference of Weblogs and Social Media*, 2009.

[6] N. Agarwal, M. Galan, H. Liu, and S. Subramanya., "Wiscoll: Collective wisdom based blog clustering," *Journal of Information Science: Special Issue on Collective Intelligence*, vol. http://dx.doi.org/10.1016/j.ins.2009.07.010, 2009.

[7] B. Kuo, T. Hentrich, B. Good, and M. Wilkinson, "Tag clouds for summarizing web search results," in *WWW*, 2007.

[8] M. Porter, "An algorithm for suffix stripping," *Program: electronic library and information systems*, vol. 40, no. 3, pp. 211–218, 2006.

[9] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *JAIS*, vol. 41, no. 6, pp. 391–407, 1990.

[10] N. Agarwal, H. Liu, L. Tang, and P. Yu, "Identifying the influential bloggers in a community," in *WSDM*, 2008.

[11] J. Leskovec, L. Adamic, and B. Huberman, "The dynamics of viral marketing," *ACM Transactions on The Web (TWEB)*, vol. 1, no. 1, 2007.

[12] P. Pierson, *Politics in time: History, institutions, and social analysis*. Princeton University Press, 2004.

[13] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," in *IJCAI*, 2005.

[14] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *UAI*, 2004.

[15] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Latent friend mining from blog data," in *ICDM*, 2006.

[16] M. Schwartz and D. Wood, "Discovering shared interests using graph analysis," *Comm. of the ACM*, vol. 36, no. 8, pp. 78–89, 1993.

[17] K. Fujimura, "The eigenrumor algorithm for ranking blogs," in *WWW*, 2005.

[18] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *WWW*, 1998.

[19] G. Mishne, "Information access challenges in the blogspace," in *Proceedings of International Intelligent Information Access (IIIA)*, 2006.

[20] Y. Chen, F. Tsai, and K. Chan, "Blog search and mining in the business domain," in *The International workshop on Domain Driven Data Mining*, 2007.

[21] N. Agarwal and H. Liu, *Modeling and Data Mining in Blogosphere*, ser. Synthesis Lectures on Data Mining and Knowledge Discovery, R. Grossman, Ed. Morgan and Claypool, 2009, vol. 1.

[22] C. Brooks and N. Montanez, "Improved annotation of the blogosphere via autotagging and hierarchical clustering," in *WWW*, 2006.