Searching for "Familiar Strangers" on Blogosphere: Problems and Challenges

Nitin Agarwal*, Huan Liu*, John Salerno[+], and Philip Yu[#]
** Computer Science & Engineering, School of Computing Informatics*
Arizona State University, Tempe, AZ 85283
[+] *Air Force Research Lab/IFEA, Rome, NY 13441*
[#] *IBM T.J. Watson Research Center, Hawthorne, NY 10532*
{nitin.agarwal.2,huan.liu}@asu.edu, john.salerno@rl.af.mil, psyu@us.ibm.com

Abstract

In this work, we examine familiar strangers on Blogosphere and issues of finding them. In our daily life, familiar strangers, as coined by Stanley Milgram, do not know each other, but frequently exhibit some common patterns. Blogosphere is a part of the Web where bloggers post in individual or community blog sites. The nature of the Web is a scale-free network, which determines that a power law distribution applies to bloggers. That is, the majority bloggers are only connected with a small number of fellow bloggers, and these blogging groups are largely disconnected from each other. Familiar strangers on Blogosphere are not directly connected, but share some patterns in their blogging activities. We present a new problem: Aggregating familiar strangers on Blogosphere that allows for better personalized services, targeted marketing, exploration of new business opportunities, and predictive modeling and marketing. Finding familiar strangers on Blogosphere presents a challenge resulting from their disconnectedness. We look at typical blogs and understand the status quo, while seeking innovative ways to improve business intelligence. We define the problem of searching for familiar strangers on Blogosphere, elucidate the significance of doing so, study the challenges of finding them, and present and discuss some potential approaches.

1. Familiar Strangers on Blogosphere

The advent of Web 2.0 [2] has started a surge of open-source intelligence via online media such as blogs. Since more and more people are participating in Web 2.0 activities, it has generated enormous amounts of *collective wisdom* or *open-source intelligence.* Web 2.0 has allowed the mass not only to contribute and edit posts/articles through blogs and wikis, but also enrich the existing content by providing tags or labels, hence turning the former information consumers to the new producers. Allowing the mass to contribute or edit has also increased collaboration among the people unlike Web 1.0 where the access to the content was limited to a chosen few. Blogs are invigorating this process by encouraging the mass to document their ideas, thoughts, opinions, views reverse chronologically, called *blog posts*, and share them with other *bloggers*. These blog posts are published on *blog sites* and the universe of these blog sites is called *Blogosphere*. Familiar strangers on Blogosphere are not directly connected, but share some patterns in their blogging activities.

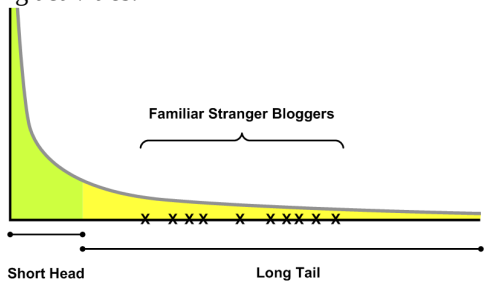

Figure 1: Familiar Stranger Bloggers and Long Tail Distribution

Blogosphere contains both single authored blog sites known as individual blog sites and multi-authored blog sites or community blog sites. In individual blog sites, only one author creates blog posts and readers are allowed to comment on these posts, but the readers cannot create new entries. In community blog sites, several authors can create blog posts and comments. Readers are allowed to comment but only registered members of the community can author blog posts. Based on these different entities on the blogosphere we have two types of familiar strangers on Blogosphere: groups and individuals. It is highly likely that both these types of familiar strangers occur in the Long Tail [1] as depicted in Figure 1, because the bloggers in the Short Head are highly authoritative which means they are highly connected, hence less chances of being strangers. Moreover, existing search engines return relevant results only from the Short Head, so it is interesting and challenging to study the ones that appear in the Long Tail. In this work we focus on individual familiar strangers.

Given a blogger *b*, we aim to find *b*'s familiar strangers, and together, they form critical mass such that (1) the understanding of one blogger gives us a sensible and representative glimpse to all, (2) more data about

familiar strangers can be collected for better customization and services (e.g., personalization and recommendation), (3) the nuances among them present new business opportunities, and (4) knowledge about them can facilitate predictive modeling and trend analysis.

Familiar strangers on Blogosphere are the niches of business opportunities. They are distributed over the blogosphere and each is in a small group. Each group is isolated and its size is also small such that the need for a zoom-in study is often ignorable. For example, it is not cost-effective to hire an expert to personally study a single blogger. However, aggregating familiar strangers can open up new opportunities. For the same example, the knowledge of one blogger's personalization can now be transferred to these familiar strangers so that the previous worthless zoom-in study becomes meaningful. In addition, their aggregation can provide a rich body of data that can be used for accurate personalization and mining for patterns. Next we study the purposes of finding familiar strangers on the blogosphere.

2. Need for Aggregating Familiar Strangers

As we know, the Web follows the distribution of a power law. Since the blogosphere is a part of the Web, the power law naturally applies to the blogosphere in the sense that except for a small percentage of blog sites, the majority of blog sites belong to the Long Tail. In particular, many bloggers are active locally with limited connections to other bloggers. Here is the dilemma: Before a blogger becomes prominent or in the Short Head, it is not worth paying particularly customized attention to the blogger; and the blogger cannot be well targeted for otherwise potential business opportunities (i.e., niches). As depicted in Figure 2, ads generated by Google AdSense are not relevant. A not-well connected post (in the Long Tail) does not have sufficient information for link-based approaches to find relevant ads. To do better requires a good number of bloggers that can provide more data for accurate automatic personalization for targeted marketing. Now we elaborate the impact of finding familiar strangers on the blogosphere from the perspective of marketing with Web 2.0.

The underlying concept of familiar strangers is that they share some patterns and routines (or commonalities), although they are not directly connected. Connecting them to form critical mass will not only expand a blogger's social network, but also increase participation to move from the Long Tail toward the Short Head. With Web 2.0, the new marketing 4Ps [3] are personalization, participation, peer-to-peer, and predictive modeling. Personalization is to customize products and services through the use of the Internet with emerging social media and advanced algorithms. The pervasive use of the Web technologies extends the Long Tail even longer, which makes personalization more important as well as more difficult in a cost-effective sense. It is important because impersonalized ads will

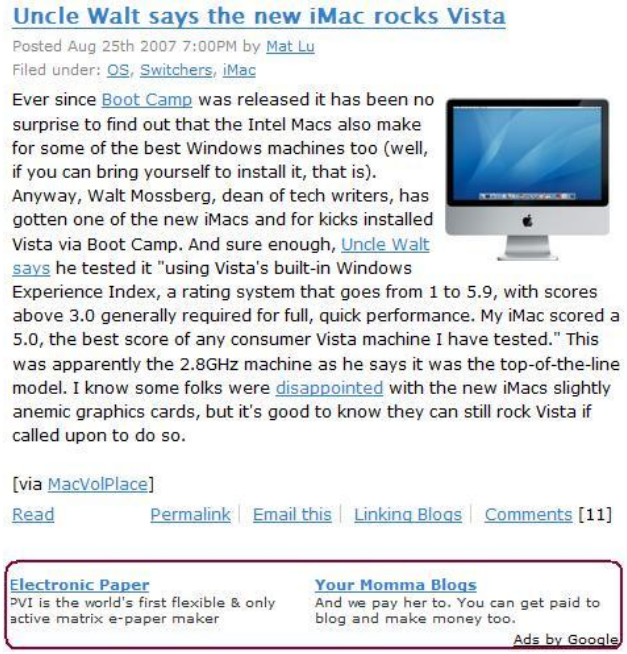

Figure 2: A blog post with irrelevant ads

likely be ignored and hence it defeats the purposes of attaching ads in the first place. It is difficult because the data is too sparse to be useful for accurate personalization. Participation allows a customer to participate in what the brand should stand for; what should be the product directions, or which ads to run. Finding familiar strangers on the blogosphere can increase the customer base of similar interests, which can encourage participation due to the crowd effect as reputation can significantly increase as the customer base expands. Reputation and expression are among major motivations for bloggers to engage in activities; and shared interests will encourage them to participate more actively. Finding familiar strangers makes peer-to-peer feasible, which refers to customer networks and communities where advocacy happens. Peers usually trust peers. Knowledge transfer or information flow among peers becomes smoother and more likely to be useful. Familiar strangers share some commonalities but can have varied deviations. The differences among them can be considered relevant niches for new business. Predictive modeling refers to employing inductive algorithms to learn predictive models from data in order to predict trends. Typical examples of predictive modeling include regression, classification, clustering, association rule mining, and other induction-based learning algorithms. Predictive modeling helps figure out what is going on or likely to happen and get ready to be among the first who act on the new business opportunities

when it does happen. Alternatively, one can proactively prepare effective measures to respond and react in shortest time possible. Without being discovered from the Long Tail, it is hard to be differentiated from other fellow members in the Long Tail, one is less likely to receive necessary attention that warrants better services.

Aggregating familiar strangers can have significant impact on moving from the Long Tail to the Short Head. However, the mere fact that they are strangers presents challenges to reach many disconnected bloggers over the blogosphere effectively and efficiently. We next examine some challenges.

3. Problem and Challenges

Having discussed the need for identifying familiar strangers on the blogosphere, we try to formally define them here. Bloggers, with their blogging behavior, tend to create social relationships with peer bloggers. However, most of them (~97%) are locally connected with limited links to other bloggers, thus in the Long Tail. The goal of this work is to aggregate familiar strangers. Given a blogger b, familiar strangers to b are a set of bloggers $B = \{b_1,b_2,\ldots,b_n\}$, who share common patterns as b, like blogging on similar topics, but have never come across each other or have never related to each other. Basically, every pair $\{b,b_j\}$ of bloggers, where $1\leq j\leq n$, blog on similar topics making them *familiar* or sharing the latent process that inspires them to do so. Similarity will be discussed in Section 4 (Similarity-based Approach). Nevertheless, $\{b,b_j\}$ still remain *strangers* because of no direct interaction between them either in terms of links in their blog posts or each one's presence in the other's social network. For the pair of $\{b,b_j\}$ to be *total strangers*, two conditions should hold true:

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

Failing one of the two conditions would make them *partial strangers*. For example, many adults in the US know of President Bush, but not vice versa. Henceforth, strangers are total strangers.

Since these familiar strangers are identified on the blogosphere, organizational differences in the blogosphere eventuate disparate types of familiar stranger bloggers. The blogosphere can have many social networking sites (MySpace, Orkut, Facebook, etc.). Each site can have many blog sites (or communities). We divide familiar strangers into three broad categories: 1. *Community-level familiar strangers* – two bloggers b_{ix} and b_{iy} of the same community C_i (as shown in Figure 3), 2. *Networking-site-level familiar strangers* – two bloggers b_{ix} and b_{jy} of different communities C_i and C_j, respectively on the same site (shown in Figure 4), and 3. *Blogosphere-level familiar strangers* – two bloggers b^m_{ix} and b^n_{jy} in two different communities C_i and C_j which are under different social networking sites, S^m and S^n, respectively (as shown in Figure 5.) Clearly, community-level familiar strangers are the easiest to find by studying one community at a social networking site. Identifying networking-site-level familiar strangers is still relatively easy. Identifying blogosphere-level familiar strangers is the most challenging. When we span across different social networking sites, we run into various problems like blogger identity mapping, related community identification, etc. The same blogger could use different identities on different social networking sites. It is challenging to make sure we are dealing with the same blogger, or two seemingly different bloggers on two different networking sites may be the same person. On the other hand, it would be exemplar to be able to find those bloggers using different identities at disparate networking sites as familiar strangers to themselves.

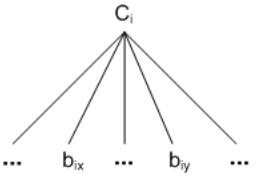

Figure 3: Community level familiar stranger bloggers

Examples of these different types of familiar strangers are: 1. *Community-level familiar stranger bloggers* – on MySpace a community called "A group for those who love history" has 38 members; two members, "Maria" and "John" blog profusely on the similar topic, but they are not in each other's social network.

2. *Networking-site level familiar strangers* – we considered two groups on MySpace, "The Samurai" and "The Japanese Sword" consisting of 32 and 84 members,

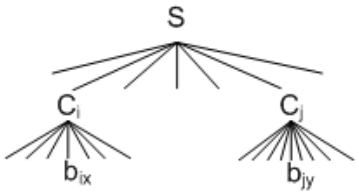

Figure 4: Network-site level familiar strangers

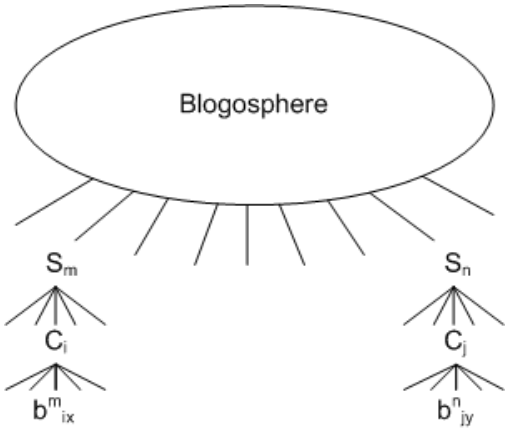

Figure 5: Blogosphere level familiar stranger blogger

respectively. These communities are located at the same networking site, and the two bloggers discuss about Japanese martial arts. We looked at the social network of the top bloggers in these groups, Marc from "The Samurai" and Jeff from "The Japanese Sword". Neither of them is in the other's social network. This implies, though being active locally and discussing on the same theme, the two bloggers are still strangers.

3. *Blogosphere-level familiar strangers* – We consider two different social networking sites, MySpace and Orkut. We manually search for similar communities on both networking sites. We picked "The Samurai" from MySpace and "Samurai Sword" from Orkut. "The Samurai" contains 32 members and "Samurai Sword" contains 29 members, which means both would lie in the Long Tail. Both groups are located in completely different networking sites. Top bloggers from the respective communities in MySpace and Orkut, "Marc" (USA) and "Anant" (India), respectively, share the blogging theme but they are not in each others' social network. The above example illustrates the existence of blogosphere-level familiar strangers.

The problem of finding familiar strangers 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 familiar stranger bloggers mentioned above. Similarity can be defined by topics, bag of words, tag clouds, etc., and will be discussed in Sec. 4.

One challenge is that a fragmented Web entails the fragmented blogosphere. Finding familiar strangers is essentially a problem of searching the Long Tail: starting from a given blogger, we want to find familiar strangers. Given that a blogger has a social network, it seems sensible to start the search with the social network. The hope is that a familiar stranger is the blogger's friend's friend (or *n*-th friend's social network), i.e., a familiar stranger of Blogger *b*'s 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 a type of naïve link analysis that entails exhaustive search. Assuming each blogger has a social network of 10 friends, the search cost is $O(10^{10})$ after exhausting bloggers who are 10 links away from the first blogger. It might very likely find familiar strangers, 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 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 familiar strangers 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. We will have to address this challenge in order to demonstrate the efficacy of various approaches and comparative findings.

4. Finding Familiar Strangers

The Long Tail phenomenon demands novel ways to find and connect little ones so that together they can emerge to become true niches as business opportunities. Given the definitions of familiar strangers, we generally have access to three types of information and data: (1) blogger *b*'s social network or *b*'s immediate links to other bloggers or posts, (2) *b*'s blog posts, and (3) blogger *b*'s context. For each type, we investigate how to leverage the existing search engines and APIs to develop algorithms such that the feasibility and potentials of the corresponding approaches can be evaluated.

The first approach is link-based. It searches for familiar strangers via a blogger's social network, or naïve link analysis. Conceptually, it can be formulated as a matrix analysis problem as follows: Given a blogger-to-blogger $d \times d$ matrix $\mathbf{A}$ of d bloggers representing pair-wise direct links, $\mathbf{A}^n$ will reveal who can be reached via *n* links. Challenges are (1) finding the link information at each step from step 1 to step *n*, and (2) making the huge, sparse matrix multiplication practical and efficient. This is basically an exhaustive search process.

The second approach is similarity-based. It searches via a blogger *b*'s posts, as illustrated in Figure 6. Intuitively, *b*'s posts, $\{p_1,p_2,\ldots,p_n\}$ contain a rich amount of information such as text, links, and tags. In absence of tags, existing approaches like [4] can be used to discover topic structures of the blog posts. One way of finding familiar strangers is to use the tag cloud [5] or topics of *b*'s post as query terms (or query expansion) to find those highly similar posts $\{q_1,q_2,\ldots,q_n\}$ by employing search engines or meta-search engines. Since we basically use the ranking functions of the search engines, the top ranked ones might miss those in the Long Tail. Top results would always have high authority or linked by several authoritative blog posts. Hence existing search engines tend to produce results from the Short Head. If we remove the top results from the search, we may end up with irrelevant results. This results in a dilemma. A first step is to evaluate the posts returned by representative search engines and meta-search engines in a controlled domain to observe how disparate types of results are distributed in terms of relevance. Once related and relevant blog posts are obtained can we study the authors of these blog posts, $\{b_1,b_2,\ldots b_n\}$ and look for familiar strangers, as defined in Section 3.

The third approach is context-based, which makes use of a blogger's context as displayed in Figure 7. The context of a blogger could be gleamed from the community he is a part of. Using the community tag

information the search for familiar strangers, and bloggers could be restricted to other communities of a similar

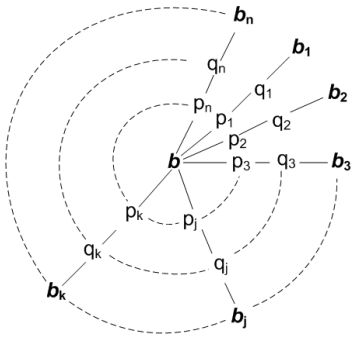

Figure 6: Searching via Blogger's Blog Post

category. The key is to use local information that can be identified through each individual to guide a directed search. Directly searching the Long Tail for familiar strangers turns this approach to the first approach – exhaustive search. The alternative is to use the tag information to determine those relevant categories that can be found in the Short Head, then lead the blogger to those reachable via some sites in the Short Head. This way, it avoids exhaustive search. First, we use the context information to find relevant sites in the Short Head. Second, we further filter those groups/bloggers at each relevant site in the Short Head that act as connectors (e.g., b_s in Figure 7) between familiar strangers in the Long Tail (e.g., *b* and *b'*). If the familiarity between blogger *b*, and his/her familiar stranger from the Long Tail, *b'* is depicted by *b~b'* then the lower bound to this similarity can be $b \sim b_s$, where b_s is the connector from the Short Head.

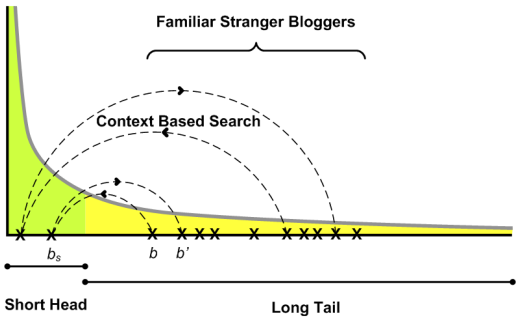

Figure 7: Searching via Blogger's Context

Some intuitive baseline methods for finding familiar strangers for comparative study are related to recommending and searching/ranking. The first is Amazon's recommendation-based approach. Each blogger could be represented with a unique profile, and bloggers can be compared based on the similarity between their profiles. But the obvious problems are (1) whether there exist databases of blogger profiles, and (2) how to construct profiles for each blogger in advance. This approach may be feasible for a limited-size domain, but will encounter a scaling-up problem when searching the blogosphere. A changing profile could cause a severe problem. Another baseline approach could be to use searching/ranking engines and obtain the blogs/bloggers similar to the one at hand. But as mentioned before, the results obtained will, more often than not, belong to the Short Head. The key issue is whether it can find those relevant ones in the Long Tail. These research issues await interesting solutions.

5. Further Study and Future Work

Based on the Power Law distribution of the blogosphere and motivated by discovering and connecting niches in the Long Tail, we present a new problem – searching for familiar strangers on the blogosphere, which is a Long Tail problem instead of a Short Head problem that can be addressed by state-of-the-art approaches. This Long tail problem raises many technical challenges. We describe the need of doing so, give working definitions, describe three types of familiar strangers, illustrate the challenges, and provide some potential solutions and baseline approaches. The three approaches employ different types of information (links, similarity, and context). Because of the nature of the Long Tail, novel solutions need to be developed. We are conducting experiments at the time of writing. Based on experimental results and findings, we will explore how to integrate the three approaches and devise an effective and efficient approach that guides the search combining link, similarity, and context in search of familiar strangers.

6. Bibliography

[1] Chris Anderson. The Long Tail: Why the Future of Business is Selling Less of More. Hyperion, 2006.

[2] Tim O'Reilly. What is web 2.0 - design patterns and business models for the next generation of software. September 2005 http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

[3] Idris Mootee. High Intensity Marketing. SA Press, 2001.

[4] J. Allan, editor. Topic Detection and Tracking: Event-based Information Organization. Kluwer, 2002.

[5] Byron Y-L Kuo, Thomas Hentrich, BenjaminGood, and Mark Wilkinson. Tag clouds for summarizing web search results. In Proceedings of WWW, 2007.