A qualitative method to find influencers using similarity-based approach in the blogosphere

Eunyoung Moon Sangki Han
Graduate School of Culture Technology
KAIST
Deajeon, Republic of Korea
{silverm913, stevehan}@kaist.ac.kr

Abstract—**Current blog systems rank 'A-list' bloggers, but they are not necessarily influential. To differentiate influential bloggers from popular bloggers, we present important groundwork for identifying influential bloggers by weighting readers based on homophily and vulnerability with bloggers. We develop the Quantifying Influence Model (QIM), which attempts to measure the influence score of bloggers. QIM is composed of two components: (1) interpersonal similarity presents the interaction among bloggers and like-minded readers, and (2) degree of information propagation represents how many readers a blogger has, where the readers diffuse the blog posts via scrapping engagements. Our study shows that weighting blog social ties can differentiate influential bloggers from popular bloggers, and what make bloggers influential or popular.**

Keywords-blogosphere;influence;popularity

I. Introduction

According to Jupiter Research in 2009, 50% of internet users search through information in the blogosphere before buying products. With the exponential growth of the blogosphere, a massive number of blog posts are being created so that readers must sift through many posts in order to get proper reviews and opinions. Accordingly, the significance of influential bloggers has attracted growing attention recently and identifying influential bloggers is much-needed in that they are potential market-movers, act as word-of-mouth advertising of several products and services, and help in customer support and troubleshooting [20].

Although current blog systems list 'A-list' bloggers according to the statistical criteria, they are not necessarily influential. This is mainly because simply commanding a high portion of traffic does not stem from the blogger's influence. Also, in the blogosphere, popularity cannot be interpreted as influence [17]. Thus, among 'A-list' bloggers, differentiating influential bloggers from popular bloggers as well as identifying influential bloggers should be addressed.

A. Background

Research on finding influencers has long been studied in the field of sociology. From this, there are underlying concepts that should be reflected in our research.

First, a fundamental property of social network is that people tend to form relationships with others who are similar to them [18]. Mcpherson et al. stated that homphily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people. Second, homophily also functions as the significant factor in the influence process. The process of social influence leads people to adopt behaviors exhibited by those they interact with [6]. This phenomenon, social selection, is manifested in many settings where new ideas diffuse by word-of-mouth or imitation through a network of people [22].

According to [24], these two concepts are also shown online. Their findings indicate that internet users seek out interactions with like-minded individuals who have similar values and thus become less likely to trust important decisions to people whose values differ from their own. Accordingly, in identifying influential bloggers, we take two important principles into consideration.

In addition to this, our study is mainly based on that large cascades of influence are driven not by influential, but by a critical mass of easily influenced individuals [25]. Thus, in this work, we reflect the propensity to be influenced in modeling.

B. General Terms Used

In the blogosphere, there are several ways in which readers can engage in the blogosphere and they are expressed as variant terms although they are the same behavior. Thus, it needs to define the terms of engagements used in this paper.

1) Comment

Comments are usually texts published by readers in the 'comment space'. They have a link for the reader's own blog.

2) Sympathy

Expressing sympathy appears in various types according to blog site, for example, in Xanga[1], 'recommendation' is used, in Digg[2], 'voting' is used so that users can give their votes in the form of a dig, and in Naver blog[3] 'sympathy' is used. In Naver, readers express their sympathy with blog posts in the form of pressing 'sympathy' button. This also has a link for the reader's own blog. Hence, although different types are used in different blog sites, these are the variants of the same behaviors. We refer this as 'sympathy'.

3) Scrap

When readers want to take blog posts, they engage by

[1] www.xanga.com
[2] www.digg.com
[3] blog.naver.com

978-0-7695-4211-9/10 $26.00 © 2010 IEEE
DOI 10.1109/SocialCom.2010.39

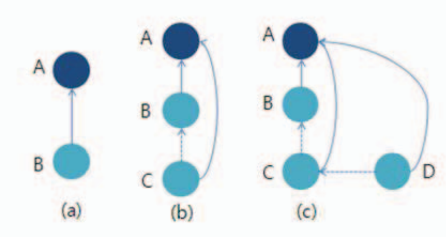


Figure 1. Cascades of blog post via scrapping

sharing blog posts and content in their own blogs or on another domain. This engagement is often known as a ‘share’. In Naver blogs, it is termed as a ‘scrap’, and a ‘scrap’ occurs by n-degree of readers. Figure 1 illustrates this statement. In Figure 1, each node represents a blog post- (a) node B creates a link to node A by scrapping, (b) node C takes the action of scrapping node B’s scrapping content that originated from node A and, in Naver blogs, node C’s scrapping link is directed to node A. In (c), this process is also conducted in the case of node D which scraps node C’s scrapping contents that originated from node A. Node D’s link is also directed to node A. In this way, ‘scrapping’ engagements continues and we refer to it as a ‘scrap’.

Note that scrapping is not copying part or all of content. It involves a link to the author’s blog, and most ‘A-list’ bloggers in Naver prevent readers from copying for their copyright. That is, to take resources readers should use a ‘scrap’.

In this paper, influencers are defined as those who have influential power to the point that one can change others thinking or behavior. Among influencers, we identify opinion leaders [3], who can do this only with the subject in which (s)he posses great knowledge. To find opinion leaders, we present the Quantifying Influence Model (QIM) and using blog data from Naver, the largest blog service in Korea, we find that the QIM can differentiate influential bloggers from popular bloggers.

II. RELATED WORK

A. Web page ranking algorithms

The two best known webpage ranking algorithms are PageRank [19] and HITS [12]. In PageRank algorithm, important web pages are linked by many link citations like in the literature, and indirect citations are also considered. HITS use an iterative algorithm to evaluate an authority weight and a hub weight for each page in a collection of related web pages. After completing computation, the web pages with top authority scores are considered as authority pages, and those with top hub scores are considered as hub pages.

However, applying webpage ranking to the blogosphere in order to rank influential bloggers is insufficient [15]. This is because measuring influence in the blogosphere is different from finding authoritative webpages for the following reasons. First, blogs in the blogosphere are sparse and the webpage ranking algorithms do not perform well. In addition, while a webpage may acquire authority over time, a blog post’s influence diminishes over time. That means the blogosphere is dynamic in a shorter time because a number of new sparsely-linked blog posts appear every day.

B. Influence models in the online social networks and blogosphere

There have been some studies on influence in online social networks. Kempe et al. [13,14] propose influence models to mathematically simulate the spread of information in social networks. They identify which nodes will maximize the spread of information. In [9], Gruhl et al. study information diffusion of various topics in the blogosphere. They propose an expectation maximization algorithm to predict the likelihood of a blogger linking to another blogger. Adar et al. [1] present the iRank to rank blogs based on informativeness. iRank finds the path of infection and blogs that initiates epidemics. Agarwal et al. [2] identify influential bloggers based on the properties of their blog posts. This study considers the characteristics of blog posts such as the novelty, the eloquence of blog posts and blog post length.

However, previous studies have some limitations as follows. First, they focus on the influential bloggers and/or their blog posts themselves. Although the distinctive properties of influential bloggers should be considered, the aspect of who is influenced and propagates blog posts has to be reflected more significantly. It is based on the study that the most social change is driven not by influentials, but by easily influenced individuals [25]. Second, they do not consider that individuals vary in their willingness to adopt new idea or products [8,23]. They treat all the blog social ties among bloggers and readers equally despite the fact that each reader has different threshold to be influenced. In this paper, we address these limitations by considering the quality of blog social ties according to their importance.

III. QUANTIFYING INFLUENCE MODEL (QIM)

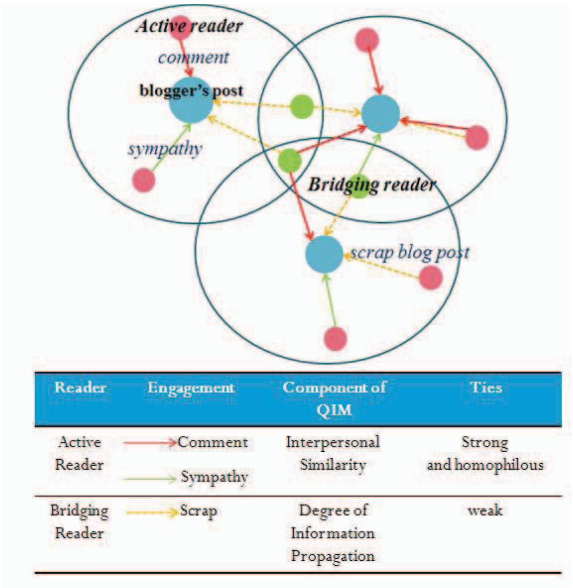
To quantify the influence score of blog posts, the Quantifying Influence Model (QIM) is proposed. The concept of QIM is shown along with a method to compute it to identify influential bloggers. We first describe observable engagement and define blog social ties as indicators of QIM.

A. Indicators of QIM

There are the various methods in which readers can engage in the blogosphere. Their engagement can be largely observed at two levels, the interpersonal level and system level.

1) Engagement at the Interpersonal level

At the interpersonal level, readers’ observable engagements with blog posts are commenting, expressing sympathy, and writing trackback. A comment is the most basic form of weblog social interaction [17]. It serves as a simple and effective means for bloggers to interact with their readership. Expressing sympathy is a way for readers to express views in a manner simpler than a comment. As readers interact with bloggers by contributing in the form of a response to specific blog posts, these engagements are observed as indicators of influence driven by readers. Finally, trackback is an automatic form of communication that occurs when one weblog references another [17]. The trackback system gives bloggers and readers an awareness of who is



Reader	Engagement	Component of QIM	Ties
Active Reader	→Comment →Sympathy	Interpersonal Similarity	Strong and homophilous
Bridging Reader	→Scrap	Degree of Information Propagation	weak

Figure 2. Concept of QIM

discussing their content outside of the original blog. Thus, trackback can be an indicator of being influenced. However, in Korea, writing trackback is rarely used by bloggers and readers. Hence, this factor is not included in the proposed model.

In brief, commenting and expressing sympathy with blog posts are considered as indicators at the interpersonal level. Readers who take part in these engagements are defined as *active readers*.

2) Engagement at the System level

By scrapping blog posts, readers send them to their own blogs or to another domain. Through the action of a reader's scrapping, blog posts can be propagated to the broader system beyond the interpersonal level. This engagement, as described in detail in section 1-(B)-(3), is driven by the n-degree of readers in Naver blogs. That is, scrapping is considered as an indicator at the system level. Readers who scrap blog posts are defined as *bridging readers*.

B. Concept of QIM

This section presents the concept of Quantifying Influence Model (QIM), as illustrated above, to measure the influence of blog posts (Figure 2), and a method to weight readers' thresholds with respect to their engagements.

As described in 3-A, influenced readers take action on two levels; thus, we define an *influential blogger* as one having both *active readers* and *bridging readers*. Moreover, as this influence is derived from both types of readers, we define influence as consisting of two components, one from active readers and the other from bridging readers.

The components of QIM are as follows: (1) *Interpersonal similarity* is derived from active readers given that *active readers* are those who have relatively strong and homophilous ties with bloggers and who have repeated interaction through commenting and/or expressing sympathy, and (2) *Degree of information propagation* is driven by bridging readers because *bridging readers* are those who have weak ties with bloggers and function as a channel of information by scrapping blog posts. In addition, to reflect the dynamics of the blogosphere, we introduce the concept of window size, which is defined as the period during which a blogger writes at least a single blog post related to a specific genre.

With these two components, QIM measures the influence of blog posts. This is expressed as follows (1):

$$Influence\ (\text{blog post}, \Delta t) = (Interpersonal\ Similarity, Degree\ of\ Information\ Propagation)\ (1)$$

QIM has metrics that distinguish it among other models in that it harnesses a qualitative method which reflects the quality of blog social ties. This is in contrast to other metrics that consider blog social ties equally. (1) It captures the importance of readers who make the flow of influence and information rather than focuses on the influential bloggers. (2) It measures both active readers' and bridging readers' weighted values according to their propensity to be influenced.

The manner of quantifying each component is then explained and the influence of blog posts is computed in the subsequent sections.

1) Interpersonal Similarity

In the field of sociology, influence at the interpersonal level occurs in that people tend to form relationships with others who are already similar to them and adopt behavior exhibited by those they interact with [18]. This phenomenon also appears in the blogosphere. In the blogosphere, bloggers tend to build relationships with other bloggers that share similar interests and tend to visit the same blogs frequently [11]. Through this process, readers interact with like-minded bloggers and are influenced by them.

Accordingly, interpersonal similarity represents the influence derived from interactions among a blogger and like-minded readers. In the online realm, homophily between two individuals is defined as a shared interest and mindset [5]; hence, to compute interpersonal similarity we harness interest similarity. The interest similarity computed between a blogger and his/her reader reflects the probability of a reader being influenced by a blogger's posts. The higher similarity between the two, the more influenced the reader will be.

To compute interest similarity, we use tag-similarity because the blogger's interests are implicitly and concisely represented by tags, subdividing their blogs into topics or themes [4,10,16]. That is, the occurrence of common tags represents their common interests. Hence, it is reasonable that tag-similarity represents online homophily among bloggers and readers. As similarity metrics, we use the Jaccard coefficient [21] to measure the level of similarity between a blogger and a reader (2).

$$\text{Jaccard coefficient}(B, R) = \frac{\text{tags}_{B\cap R}}{\text{tags}_B + \text{tags}_R - \text{tags}_{B\cap R}} \qquad (2)$$

This coefficient is calculated as the number of tags contained in a set of B's tags and R's tags and is normalized by the number of elements of the union of a collection of set of B's tags and R's tags. For instance, if B has a set of tags, {cooking, muffin, salad} and R has a set of tags, {cooking, mocha, cake, jam}, the overlapped tag is {cooking} and the

Table 1. How to determine bridging reader's weighted value

	blog post$_1$	blog post$_2$	…	blog post$_K$	…	blog post$_M$	Bridging reader's weighted value (=Total number of scrapping blog posts)
Bridging reader$_1$	A_1	A_2	…	A_K	…	A_M	$W_1 (=A_1 + A_2 + \cdots + A_K + \cdots A_M)$
Bridging reader$_2$	B_1	B_2	…	B_K	…	B_M	$W_2 (= B_1 + B_2 + \cdots + B_K + \cdots B_M)$
…	…	…	…	…	…	…	…
Bridging reader$_J$	J_1	J_2	…	J_K	…	J_M	$W_J (= J_1 + J_2 + \cdots + J_K + \cdots J_M)$
…						…	…
Bridging reader$_N$	N_1	N_2	…	N_K	…	N_M	$W_N (= N_1 + N_2 + \cdots + N_K + \cdots N_M)$
Total							$\sum_{i=1}^{N} W_i$

elements of the union of the two sets, B and R are {cooking, muffin, salad, mocha, cake, jam}. Thus, in this case, the value of the Jaccard coefficient is 1/6=0.167. That is, if no overlapping tags exist between a blogger and a reader, the coefficient has a value of 0. If all tags are shared between the two, it has a value of 1. However, if this model is applied to a domain where controversy exists, it requires semantic processing to distinguish whether the meaning of tagging is positive or negative.

The formula for interpersonal similarity is formulated as follows (3): For blog post$_k$, interpersonal similarity is computed by the tag-similarity sum among a blogger and his/her readers who comment and express sympathy with blog post$_k$.

$$Interpersonal\ similarity\ (blog\ post_k) \\ = \sum \mathrm{Jaccard}\ (\mathrm{B}, \mathrm{R}_{\mathrm{comment}}) + \sum \mathrm{Jaccard}\ \left(\mathrm{B}, \mathrm{R}_{\mathrm{sympathy}}\right) \quad (3)$$

In formula (3), B denotes the blogger, $R_{comment}$ is a reader who leaves a comment on B's post, and $R_{sympathy}$ is a reader who expresses sympathy with B's post. Here, these abbreviations are used for these two types of readers, $R_{comment}$ and $R_{sympathy}$.

2) Degree of Information propagation

Information propagation is a phenomenon in which an action or idea becomes widely adopted due to the influence of others, typically, neighbors in some network [8]. In the blogosphere, information propagation occurs through bridging readers' scrapping engagements, thus, blog posts are propagated from one segment to another segment.

However, each bridging reader has a different tendency to take the action of scrapping, as each one has a distinct propensity to scrap blog posts. For this reason, differentiating each bridging reader is necessary rather than simply considering each one as equal. To facilitate this, we define the degree of information propagation as the degree of spreading by weighted bridging readers. The following paragraphs describe how these readers are weighted.

Intuitively, the propagation of blog posts occurs by the process, as described in detail in 1-(B)-(3). More precisely, we describe this process as follows.

(1) Assume that there are M blog posts-{blog post$_1$, blog post$_2$,...,blog post$_M$}.

(2) *Bridging readers* are then defined as those who scrap more than one blog post among M posts.

(3) Next, for each bridging reader, the total number of scrapping blog posts is determined, as each one has a distinct propensity to scrap.

Table 1 details how bridging readers are weighted. In table 1, J_1 denotes the number of times bridging reader$_J$ scrapped a blog post$_1$, J_K denotes the number of times blog post$_K$ was scrapped by that reader, and W_J denotes the total number of times that this blog post was scrapped by bridging reader$_J$ among M posts. This follows with bridging reader$_1$, bridging reader$_2$,…,to bridging reader$_N$. That is, a bridging reader's weighted value is calculated as the total number of his/her scrapping blog posts. For example, assuming that readers who scrap blog post$_K$ are Bridging reader$_I$, Bridging reader$_J$, and Bridging reader$_K$, in such a case, the degree of information propagation of blog post$_K$ is calculated as follows (4) (see Table 1).

$$degree\ of\ information\ propagation(blog\ post_k) \\ = \frac{\mathrm{W_I} + \mathrm{W_J} + \mathrm{W_K}}{\sum_{i=1}^{N} \mathrm{W_i}} \quad (4)$$

Accordingly, for blog post$_k$, we derive the formula for the degree of information propagation as follows (5):

$$degree\ of\ information\ propagation(blog\ post_k) \\ = \frac{\sum \text{weighted value of reader who scrapped blog post}_{\mathrm{K}}}{\sum \text{bridging reader's weighted value}} \quad (5)$$

This formula (5) gives the weighted ratio of infected bridging readers.

C. Computation of QIM

To combine two components, it is necessary to determine initially which component functions as a more important factor when a reader is influenced by a blog post. Thus, the weight of each component is needed. In this work, the weights were empirically set on the basis of the results of an online survey of readers. These weights are not fixed but are flexible according to the different domains and topics. The method of setting the weight is described in 4-(B). The completed formula of the QIM is expressed below (6) and $\Gamma_1$and Γ_2 denote the weights.

$$Influence(\ \text{blog post}_{\mathrm{K}}, \Delta \mathrm{t}) \\ = \Gamma_1 \cdot \text{Interpersonal Similarity}(\text{blog post}_{\mathrm{K}}) + \\ \Gamma_2 \cdot \text{Degree of information propagation}(\text{blog post}_{\mathrm{K}}) \\ = \Gamma_1 \cdot \{ \sum \mathrm{Jaccard}\ (\mathrm{B}, \mathrm{R}_{\mathrm{comment}}) + \sum \mathrm{Jaccard}\ \left(\mathrm{B}, \mathrm{R}_{\mathrm{sympathy}}\right) \} + \\ \Gamma_2 \cdot \frac{\sum \text{weighted value of reader who scrap blog post}_{\mathrm{K}}}{\sum \text{weighted value of bridging reader}} \quad (6)$$

The process of ranking the influence score is described based on formula (6), for the blog post$_K$.

(1) First, computing each component and combining them gives, the influence score of blog post$_K$, referred to as the *iscore* of blog post$_K$. It is expressed as an *iscore(blog post$_K$)*.

(2) Then, all M blog posts have their own iscore, $\{\mathrm{iscore}(\mathrm{blogpost}_1),\mathrm{iscore}(\mathrm{blogpost}_2),\ldots,\mathrm{iscore}(\mathrm{blog\ post}_M)\}$. This set is termed I.

(3) Elements of set I are ranked in descending order, giving the top 100 iscores. We define blog posts with the top 100 iscores as *influential blog posts*.

(4) A blogger who has any blog post in the top 100 iscores is then defined as an *influential blogger*.

Accordingly, the means of identifying an influential blogger is to check if the blogger has any influential blog posts, i.e. *A blogger is influential if (s)he has more than one influential blog post*.

The method of computing the influence of influential bloggers and ranking them is then described. For *blogger$_i$* who has more than one influential blog post, his/her influence is computed as the average of iscore of his/her influential blog posts in the top 100. We express this as *Influence(blogger$_i$)* (7).

$$\begin{aligned}&\mathit{Influence}(\mathrm{blogger_i})\\&= \text{average }(\mathit{iscore}\text{ of blogger}_\mathrm{i}'\text{s influential blog post}_\mathrm{K})\\&\quad ,\text{where } 1 \le \mathrm{K} \le \mathrm{M}\end{aligned} \tag{7}$$

The influence of influential bloggers can thus be determined. Therefore, we can rank *Influence(blogger$_i$)* in a descending order and obtain the rank of influential bloggers. From this rank, the top *k* bloggers are defined as the most influential bloggers. Setting k is a challenging issue, and determining the threshold requires further research. Hence, this is done on the basis of the pertinent study described in section 4-(C) below.

IV. EMPIRICAL STUDY

A. *Naver blog and Popularity model*

To evaluate the QIM, the Naver blog was used. This is the largest blog service in South Korea, with a market share of over 70%, compared to 2% of Google[4]. The Naver blog is not only by far the most popular, but also lists 'topic-centered and A-list' bloggers on 31 topics. Accordingly, it provides an excellent opportunity to observe opinion leaders who are more involved in their main interest.

Among 31 topics, we chose the cooking domain which is the highly active and where controversy among bloggers and readers does not exist. Hence, it provides us with an opportune chance to observe bloggers and readers compared to other domains. It is possible in this domain to regard overlapped tags between a blogger and his/her reader as a degree of common interests (see 6-B). Moreover, 99 'cooking-centered and A-list' bloggers are considered as seed nodes, as they provide a greater opportunity to observe interaction among bloggers and readers than do non-A-list bloggers.

To compare popularity with influence, defining a popularity model is required. Based on the concept of degree from graph theory, indegree is interpreted as a form of popularity, the majority of blog platforms such as Technorati[5], Xanga, and Naver rank A-list bloggers by the simple summation of the indegree of each node. Accordingly, we define the indegree of seed nodes as the criteria of popularity. In the Naver blog, observable engagements for indegree are commenting, expressing sympathy, and scrapping. Thus, we define the popularity model as the simple summation of these three engagements. It is therefore clear that although popularity and influence are considered as the same indicators, weighting blog social ties according to the importance of a node would be the key in differentiating influence from popularity.

To find popular bloggers, we computed the popularity score of blog posts and ranked the popularity score in descending order. Thus, blog posts on the top 100 popularity score are defined as *popular blog posts* and a blogger who has any blog post on the top 100 list is defined as a *popular blogger*. Next, the popularity of a popular blogger is computed as the average of the popularity score of his/her popular blog posts in the top 100, giving the popular blogger's rank by the ranking of the average value.

B. *Data collection*

We conducted an evaluation of the QIM three times. To reflect the dynamics of the blogosphere, the window size was set according to the period of the seed nodes, from the 31st of July to the 19th of August 2009, from the 10th to the 30th of September of 2009, and from the 1st to the 20th of October of 2009. In total, 1,658 blog posts of 99 seed nodes, 55,136 comments, 26,233 engagements of sympathy, and 163,412 engagements of scrapping were collected.

Additionally, to fix two weights, an online survey of readers on the cooking domain was conducted. Considering that influenced individuals create the flow of influence [25], when choosing a recipe, the degree to which they consider the component of the QIM was assessed. Thus, the respondents were asked how much they considered the component 'taste-similarity' and the 'degree of being scrapped by many' when they choose a recipe.

156 of 991 readers rated how important they think each component is on a five-point scale. If the component is of very little importance, they should give it as a score of 1; in contrast, if they think a component is of much greater importance, they should score it as 5. Table 2 shows the results. Among the 156 respondents, 6 respondents responded to only the first question and did not answer the second question. From the result, by computing the ratio between the two average values, weight Γ_1 was determined as 1.7137 while Γ_2 as 1.

Table 2. Results of online survey to fix two weights

	Taste- similarity			Degree of being scrapped by many		
	Sum	Respondents	Avg	Sum	Respondents	Avg
Total	701	156	4.49	394	150	2.62

[4] www.google.com

[5] www.technorati.com

Table 3. Top 3 influential/popular bloggers

Rank	Number of influential blog posts	Total Influence score	Number of popular blog posts	Total Popularity score
1st	42	19.8338	1	2855
2nd	16	19.3391	10	2770.214
3rd	22	11.6100	6	1963.25

C. Result of experiment

In each experiment, we listed the top 100 influential blog posts and the top 100 popular blog posts. From the two lists of top 100 blog posts, the top influential bloggers and most popular bloggers were computed. By integrating data from three runs of experiments, the top 35 influential bloggers and the top 40 popular bloggers were determined. Interestingly, 20 bloggers were present in both of the lists; and this set of bloggers is considered to comprise those who are both influential and popular. Also, to identify the most influential bloggers and popular bloggers, the value k was set to 3 because an influential blogger is defined as an individual in the top 10% of the influence distribution [25]. Thus, tracking the top 3 influential bloggers is reasonable (Table 3).

D. A closer look at influential /popular bloggers

1) Correlation between Influence and Popularity

Before validation, to measure the strength of the association between two rank sets, QIM-based influential bloggers and popular bloggers, we used the Spearman's rank correlation coefficient. The closer the coefficient is to +1 or -1, the stronger the likely correlation. A perfect positive correlation is +1 and a perfect negative correlation is -1.

We computed the Spearman's rank correlation coefficient for the top 50th percentile as well as for the entire set of bloggers. The results were -0.303 and 0.045, respectively. This clearly shows that in the blogosphere, influence and popularity have little correlation with each other; thus, it is clear that popular bloggers are not necessarily influential bloggers, and vice versa.

2) Qualitative indicators from the most influential/popular bloggers

Also observed were two significant indicators. The readers of the most influential/popular bloggers showed distinct differences from the activity status and the blog social ties of other influential bloggers. We investigated the top 3 influential bloggers and the top 3 popular bloggers based on two indicators: (1) how many non-active readers the influential/popular bloggers have, and (2) whether or not bloggers have blog social ties with other influential bloggers.

Indicator (1) denotes that $R_{comments}$ and/or $R_{sympathy}$ closed his/her blog site. Although a reader had his/her own blog site address, all of the menus of the blog were set as non-active so that the web pages of the blog could not be viewed. We refer to these as *non-active readers*. This is important because non- active readers cannot diffuse blog posts, in this regard, they are akin to readers who only take resources to satisfy their needs.

Table 4. Number of Non-active readers

	Top 3 Influential bloggers		Top 3 Popular bloggers	
	Non-active $R_{comment}$	Non-active $R_{sympathy}$	Non-active $R_{comment}$	Non-active $R_{sympathy}$
Total	64	5	215	28
Per post	0.8	0.0625	12.647	1.6471

Table 5. Blog social ties with other 3 types of bloggers

	Influential bloggers		Influential & Popular bloggers		Popular bloggers	
	Comment	Sympathy	Comment	Sympathy	Comment	Sympathy
Top 3 Influential bloggers	12	12	7	7	0	0
Top 3 Popular bloggers	2	2	1	0	1	2

The action of readers engaging in scrapping was not observed because they already disseminated blog posts by taking the blog posts to their blogs or other domains. Thus, this observation presents whether or not edges directed into a vertex function as a channel to activate the flow of influence. Table 4 shows that the top 3 influential bloggers have 0.8 non-active $R_{comment}$ per post and 0.0625 non-active $R_{sympathy}$ per post, whereas the top 3 popular bloggers have 12.647 non-active $R_{comment}$ per post and 1.647 non-active $R_{sympathy}$ per post. Moreover, the top 3 influential bloggers in total have far fewer non-active readers than the top 3 popular bloggers. This clearly shows that a high number of comments, sympathy remarks, and actions of being scrapped do not necessarily mean influence. Therefore they are not meaningful statistics. In this sense, weighting blog social ties according to the reader's importance is very significant.

Indicator (2) denotes that the more influential blogger is, the more that blogger will receive comments and/or sympathy from other influential bloggers, similarly, important web pages become more important, more linked by other important ones in PageRank and HITS algorithm.

Table 5 shows that the top 3 influential bloggers receive more comments and sympathy from other influential bloggers, whereas the top 3 popular bloggers receive few comments and sympathy from other popular bloggers. This implies that QIM captures the importance of readers that point to a given blogger by blog social ties. Thus, bloggers have more influence if other influential bloggers point to him/her, than if some non-influential bloggers point to him/her.

V. VALIDATION

As shown in a recent study [2], there is no training and testing data for us to show the efficacy of the proposed model. That is, ground truth about influential bloggers does not exist. However, by using a reasonable reference point, we can observe tangible differences and show that the influence of influential bloggers is stronger than that of popular bloggers.

A. Criteria for Validation

In this paper, *influencers* are defined as those who have influential power and can make others change their thinking or behavior (see section 1). Thus, we validate this by showing that influential bloggers make more readers' thinking or behavior change compared to popular bloggers. To be specific, we observe that readers imitate blogger's recipes in actuality, accept his/her ideas, and personally write about the process of trying the recipe at their own blogs. This is clearly different from scrapping posts, which occurs automatically with a single click of a mouse.

Moreover, to determine readers' referral behaviors more accurately, we sought cases in which readers specifically stated that they mimicked the recipe. To identify this clearly, we used two criteria: (1) referring to a blogger's name, and (2) making link citations.

B. *Set up for Validation*

For criteria (1), to filter cases of referring to a different blogger with the same name, we investigated whether there were bloggers with the same name as influential/popular bloggers. It was found that no such bloggers existed in the top 3 influential/popular bloggers in the cooking domain.

Conducting validation based on two criteria, we checked the keywords and tags within readers' blog posts to determine that whether readers wrote about imitating a blogger's recipe.

Additionally, for verification, among the three types of readers-$R_{comment}$, $R_{sympathy}$, and R_{scrap}-the third type of reader was not included as weighting scrapping readers has a strong correlation with the simple summation of the number of scrapping actions with a correlation coefficient 0.91. In contrast, there is a very weak correlation between weighting readers who comment and those who express sympathy and a simple summation of the number of comments and sympathy expressions, with the correlation coefficient of 0.1.

C. *Result of Validation*

As a result (Table 6), we found that the top 3 influential bloggers in total had 1393 $R_{comment}$ and 650 $R_{sympathy}$, moreover, the top 3 popular bloggers had 1726 $R_{comment}$ and 349 $R_{sympathy}$. From all of these readers, the top 3 influential bloggers' $R_{comment}$ wrote 1092 posts referring to an influential blogger's name and 304 posts making link citations to the influential blogger's posts. In the case of $R_{sympathy}$, they created 928 posts referring to an influential blogger's name and 271 posts making link citations.

However, the top 3 popular bloggers' $R_{comment}$ wrote 122 posts referring to a popular blogger's name and 13 posts making link citations of popular blogger's posts, and $R_{sympathy}$ created 93 posts referring popular blogger's name and 27 posts making link citations.

Thus, it is clear that popular bloggers fail to make a change in readers' thinking or behavior and cannot trigger them into a referral behavior. From this, we can see that QIM can differentiate influential bloggers from popular bloggers.

VI. DISCUSSION

A. *Influential bloggers and Popular bloggers*

In identifying influential bloggers, the factors that make bloggers influential or popular could be determined.

First, weak ties play an important role in gaining influence or popular. Without weak ties (i.e., bridging readers), both influential bloggers' posts and popular bloggers' posts will not be propagated across the group. That is, as Granovetter [7] found in the field of sociology, weak ties also function as a bridge in the blogosphere and bloggers can gain influence or popularity based on bridging readers to some degree.

Second, we found that strong and homophilous ties make bloggers influential for two reasons. (1) QIM-based influential bloggers have far more active readers who have a relatively high degree of common interests with bloggers, constantly interact with them and consequently imitate recipes and engage in referral behaviors. (2) In addition, QIM-based influential bloggers are closely connected with each other. That is, they have active readers who are also QIM-based influential bloggers. Moreover, they refer to each other on their own blogs. As a result, they become more influential because they have more easily influenced active readers that they influence directly. This is contrast to the finding that popular bloggers have a majority of non-active readers and lurkers. Hence, differentiating influential bloggers from popular bloggers is derived from weighting active readers by tag-similarity.

Finally, we found that in contrast to influence, popularity can be obtained spontaneously or accidently. We observed that (1) popular bloggers consistently write about a wide range of 'hot topics' to foster high traffic, and (2) they have few popular posts on a cooking topic (Table 3). These properties of popular bloggers infer that they have few steady readers with common interests about a recipe and have far more lurkers, as shown by qualitative indicators and in the validation assessment. In contrast, influential bloggers are more dedicated to cooking topics; therefore, they have far more influential posts (Table 3). As a result, they have more steady readers with common interests, regularly interacting with the bloggers.

All in all, we can derive that the function of bridging readers is limited to the flow of information in order to satisfy readers' needs, whereas the function of active readers is more salient to the flow of influence.

B. *Limitations of this study*

This work has a number of limitations in that we conducted the experiments only in the cooking domain, where no controversy exists among bloggers and readers. This is mainly because bloggers mostly write about their recipes or knowledge of ingredients. For this reason, all of the overlapped tags among bloggers and readers can be considered as the degree of common interests. Thus, it can be interpreted the higher the similar coefficient between the two, the more influenced readers will be. However, if research is conducted in another domain, such as in politics or IT, the overlapping tags among them may not yield any similarity. In such a case, semantic processing would be required.

Moreover, we conducted this study in Naver with commenting, expressing sympathy, and scrapping. However, our study is not limited to this blog platform. We can apply a QIM to other blog platforms. 'Recommendation' would be a

Table 6. Result of Validation

	Rank	Number of $R_{comment}$	Number of posts referring blogger	Number of posts of citations	Number of $R_{sympathy}$	Number of posts referring blogger	Number of posts of citations
Top 3 Influential bloggers	1	854	571	161	385	411	124
	2	343	408	114	177	399	116
	3	196	113	29	88	118	31
	Total	1393	1092	304	650	928	271
Top 3 Popular bloggers	1	457	19	2	69	13	8
	2	742	47	4	158	3	0
	3	527	56	7	122	77	19
	Total	1726	122	13	349	93	27

substitute for sympathy, and ‘sharing’ can substitute for scrapping. Sharing may require some modification or other variants as well. However, it is clear that QIM is generally applicable to any blog system with some modifications.

VII. Conclusion

This paper showed that details pertaining to who influences whom as well as how, and what make bloggers influential or popular. This study has important contributions in that (1) we attempted to position this work at the intersection of social media analysis and the social science, sociology. To be specific, we reflected the concept of homophily, social selection, and reader’s threshold. By doing so, (2) we proposed a new metric, QIM, which reflects the quality of blog social ties as well as the quantity of blog social ties. QIM shows how to quantify blog social ties according to their importance. Finally, (3) by QIM, we can differentiate influential bloggers from popular bloggers. Among ‘A-list’ bloggers, who is influential or popular can be clearly determined with QIM. However, it is a start of building such a metric and we plan to apply QIM to other domains and blog platforms. We conjecture that QIM can be a useful feature in differentiating influential bloggers from popular bloggers.

In conclusion, the proposed QIM, developed by an interdisciplinary approach, holds promise in that it lays the foundation stone of a qualitative approach to finding influential bloggers.

Acknowledgment

We thank NHN for providing us with blog data and appreciate YoungJin Kwon, Ph.D candidate in computer science department of KAIST and Meeyoung Cha, researcher at Max Planck Institute for Software Systems for reviewing and giving feedback on this paper.

References

[1] Adar,E., Zhang,L., Adamic,L.A., and Lukose,R.M. Implicit structure and the dynamics of blogspace. In workshop on the weblogging Ecosystem, New York, NY, USA, May 2004

[2] Agarwal.N., Liu.H., Tang.L., Yu.P.S. Identifying the influential bloggers in a Community, Proceedings of the First ACM International Conference on Web Search and Data Mining, 2008

[3] Blackwell,R.D.,Miniard,P.,Engel,J. Consumer behavior, Mason: South-Western, 2001

[4] Brooks,C.H., and Montanez,N., An analysis of the effectiveness of tagging in blogs. In 2005 AAAI Spring symposium on Computational Approaches to Analyzing Weblogs. AAAI, March 2005

[5] Brown,J., Broderick,A.J., and Lee,N. Word of Mouth communication within online communities: Conceptualizing the online social network, Journal of Interactive Marketing Volume 21 No.3, 2007

[6] Friedkin,N.E. A structural theory of social influence, Cambridge University Press, 1998

[7] Granovetter,M.,The strength of weak ties, American Journal of Sociology, 78(May), 1360-1380

[8] Granovetter,M.,Threshold models of collective behavior, American Journal of Sociology 83: 1420-1443, 1978

[9] Gruhl,D., Guha,R., Liben-Nowell,D. and Tomkins,A. Information Diffusion through blogspace. International World Wide Web Conference, pages 491-501, 2004

[10] Hayes,C. and Avesani.P., Using tags and clustering to identify topic-relevant blogs, ICWSM, 2007

[11] Herring,S.C., Kouper,I., Paolillo,J.C., and Scheidt,L.A. Conversations in the Blogosphere: An analysis "From the Bottom Up". Proceedings of the Thirty-Eighth Hawai'I International Conference on System Sciences (HICSS-38), 2005

[12] Kleinberg.J., Authoritative sources in a hyperlinked environment. In 9th ACM-SIAM Symposium on Discrete Algorithms, 1998

[13] Kempe,D., Kleinberg,J.M., and Tardos,E. Influential nodes in a diffusion model for social networks. In ICALP, 2005

[14] Kempe,D., Kleinberg,J.M., and Tardos,E. Maximizing the spread of influence through a social network. In KDD, pages 137-146, 2003

[15] Kritikopoulos,A., Sideri,M. and Varlamis,I. Blogrank: ranking weblogs based on connecting and similarity features. In AAA-IDEA' 06: Proceedings of the 2nd International Workshop on Advanced architecture and algorithms for internet delivery and applications, page8, 2006

[16] Li,X., Guo.L, and Zhao,Y.E. Tag-based Social Interest Discovery, International World Wide Web Conference, 2008

[17] Marlow,C., Audience, structure, and authority in the weblog community. International Communication Association Conference, May 27-June 1, New Orleans, LA

[18] McPherson.M., Smith-Lovin.L., and Cook.J.Birds of a feather: Homophily in social networks.Annual Review of Sociology, 27, 2001

[19] Page,L., Brin,S., Motowani,R., and Winograd.T., The Pagerank citation ranking:Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998

[20] Scooble,R. and Israel,S., Naked conversations:How blogs are changing the way business talk to their customers, 2006

[21] Sternitzke,C. and Bergmann,I., Similarity measures for document mappin: A comparative study on the level of an individual scientist, Scientometrics Vol.78,No.1, 2009

[22] Strang,D.,and Soule,S. Diffusion in organizations and social movements. Annual Review of Sociology. 1998

[23] Valente.T.W., Social network thresholds in the diffusion of innovations. Social Networks 18 pp69-89. 1996

[24] Van Alstyne,M., and Brynjolfsson,E., Global Village or CyberBalkans: Modeling and Measuring Integration of Electronic Communities. Management Science Vol.51, Issues 6. 2005

[25] Watts. D.J., and Dodds. P.S. Influentials, Networks, and Public Opinion Formation, Journal of Consumer Research, Uchicago Press, 2007