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

Eunyoung Moon* and Sangki Han

Graduate School of Culture Technology,
KAIST, 373-1 Gusung-Dong, Yusung-Gu,
Daejeon 305-750, Republic of Korea
E-mail: silverm913@kaist.ac.kr
E-mail: stevehan@kaist.ac.kr
*Corresponding author

Abstract: With huge popularity, blogosphere has become a significant channel of information propagation. Accordingly, internet users search for information in the blogosphere and read blog posts before making decision and/or buying brand-new products. Although current blog systems rank 'A-List' bloggers, 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
2 degree of information propagation represents how many readers a blogger has, where the readers diffuse the blog posts via scrapping engagements.

To evaluate QIM, we conduct experiments with empirical data from the Naver blog, the largest blog service in Korea. Our study shows that weighting blog social ties can differentiate influential bloggers from popular bloggers, and what make bloggers influential or popular.

Keywords: influencer; popularity; homophily; blogosphere.

Reference to this paper should be made as follows: Moon, E. and Han, S. (2011) 'A qualitative method to find influencers using similarity-based approach in the blogosphere', *Int. J. Social Computing and Cyber-Physical Systems*, Vol. 1, No. 1, pp.56–78.

Biographical notes: Eunyoung Moon received her Master of Science in Culture Technology at KAIST in 2010. Her interests are topic modelling and developing mathematical models on the social media.

Sangki Han is the Head of the Institute of Social Computing. His interests are influence model, social recommendation and analytic tools for social media.

1 Introduction

The blogosphere is one of the most popular social media, and has experienced the exponential growth. At present, a massive number of blog posts are being created. For example, as of June 2010, Naver[1] blog, which is the largest blog service in Korea, has 19 million active blogs with approximately 13,000 new blogs being created every day. Such huge popularity illustrates what an important medium of information blogs have become, particularly because bloggers convey their experiences and views on topics in a more conversational manner than can be typically be found in the mainstream media. According to a 2009 study conducted by Jupiter Research, 50% of internet users search through information in the blogosphere before buying products (Qualman, 2009). Before making decisions or buying brand-new products, internet users search for information in the blogosphere and read various blog posts. However, this often requires them to sift through many unrelated or irrelevant posts in order to find the views and opinions that they are seeking. Accordingly, the significance of influential bloggers has recently begun to attract more attention. Many companies now recognise the importance of influential bloggers, who can act as potential market-movers, provide word-of-mouth (WOM) advertising for products and services, help in customer support and troubleshooting, and change the buying decisions of their readers and fellow bloggers (Scooble and Israel, 2006). Such influence encourages marketing strategies which target specific key individuals rather than the market as a whole.

Although current blog systems often use statistical criteria to identify ‘A-List’ bloggers, such labels do not necessarily reflect how influential a blogger is. Many ‘A-List’ bloggers have discovered that it is possible to temporarily obtain large amounts of traffic by writing posts related to a wide range of hot topics, as opposed to influential bloggers, who usually collect information and provide opinions on specific topics. Thus, in the blogosphere, popularity cannot be interpreted as influence (Marlow, 2004). For these reasons, the need to identify influential bloggers and differentiate them from popular bloggers should be addressed.

1.1 Background

Research on influence has long been conducted in the field of sociology. From this, there are underlying concepts that should be reflected in our work.

First, a fundamental property of social networks is that people tend to have attributes similar to those of their friends. Mcpherson et al. (2001) defined that homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people. Also, social scientists have long observed that similarity breeds connection (Bott, 1928; Huston and Levinger, 1978; Wellman, 1929).

Second, homophily also functions as the significant factor in the influence process. Friedkin (1998) studied that the process of social influence leads people to adopt behaviours exhibited by those they interact with. This phenomenon, social influence, is manifested in many settings where new ideas diffuse by WOM or imitating through a network people (Strang and Soule, 1998).

These two concepts in the field of sociology are also shown online. According to Van Alstyne and Brynjolfsson (2005), internet users seek out interactions with like-minded individuals who have similar values and thus become less likely to trust important decision to people whose values differ from their own. Not only but this work, the interplay between similarity and social ties in online communities was found (Crandall et al., 2008). Accordingly, in this work, we identify influential bloggers with two disciplines in the field of sociology: social selection and social influence.

1.2 Definition of influence

There are a number of theories and definitions on influence. Traditionally, ideas about diffusion have centred on a select group of individuals called 'influentials', who influence an exceptional number of their peers. According to the two-step theory, influentials are a small minority of opinion leaders (Katz and Lazarsfeld, 1955). Gladwell (2000) classified influentials into three types: mavens, connectors, and salespeople.

Contrary to this traditional view which emphasises the role of influentials, some have put forth the theory that most social change is driven not by influentials, but by easily influenced individuals influencing each other (Watts and Dodds, 2007). According to this theory, average individuals play significant roles in large cascades of influence, as shown by a series of computer simulations of interpersonal influence processes. These simulations quantified the relative importance of influentials by comparing the average size of a cascade initiated by an influential to that started by an average member of the population.

This study determined that diffusion of innovation depends primarily on the susceptibility of influenced individuals. This result does not underestimate the role of influentials, but simply places greater emphasis on the critical role of easily influenced individuals.

Taken both theories into account, we considered an individual's propensity to be influenced by developing a new metric for measuring and defining the influence of bloggers. In this paper, 'influencers' are defined as those who have influential power to the point that they can change others' thinking or behaviour. Among influencers, we identify *opinion leaders* (Blackwell et al., 2001), who can change other's ideas or actions only within subjects in which (s)he possesses superior knowledge. An *influential blogger* is then defined as a blogger who has influential blog posts with leading views and information on a specific genre, whose readers have been shown to follow the opinions and, diffuse the blog posts.

In order to identify influential bloggers as opinion leaders, we present the quantifying influence model (QIM). By using blog data from the Naver blog, we find that the QIM can differentiate influential bloggers from popular bloggers.

2 Related work

2.1 Web page ranking algorithms

The two best known web page ranking algorithms are PageRank (Page et al., 1998) and hyperlink-induced topic search (HITS) (Kleinberg, 1998). PageRank algorithm is a global ranking scheme, which tries to capture the notion of importance of a page. In PageRank algorithm, important web pages are linked by many link citations like in the literature, and indirect citations are also considered. Another ranking scheme, HITS assigns a hub and authority score. HITS uses 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.

Although these are the best web page ranking algorithms, applying web page ranking to the blogosphere is insufficient in ranking influential bloggers (Kritikopoulos et al., 2006). This is because ranking influential bloggers is different from finding authoritative web pages for the following reasons. First, blogs in the blogosphere are sparse so that web page ranking algorithms do not perform well. In addition, while a web page may acquire authority over time, a blog post's influence diminishes over time. That means the blogosphere is dynamic in a shorter time since a number of new sparsely-linked blog posts appear every day.

2.2 Influence model in online social networks and blogosphere

Studies on influence in online social networks and blogosphere have been conducted. Kempe et al. (2005, 2003) propose influence models to mathematically simulate the spread of information in social networks. They identify which nodes will maximise the spread of information. Gruhl et al. (2004) study information diffusion of various topics in the blogosphere. They present an expectation maximisation algorithm to predict the likelihood of a blogger linking to another blogger. Adar et al. (2004) propose the iRank to rank blogs based on informativeness. iRank finds the path of infection and blogs that initiates epidemics. Agarwal et al. (2008) present an algorithm to compute the influence of each blogger. This study considers the characteristics of blog posts such as the novelty, the eloquence of blog posts and blog post length from The Unofficial Apple Weblog which is a multi-authored blog. However, these previous studies have some limitations for the following reasons.

First, they focus on the influential bloggers and/or their blog posts themselves. Although the influential properties should be considered, the aspect of who adopts influential's view and propagates blog posts has to be reflected more significantly.

This is shown by computer simulation of 'influential hypothesis' (Watts and Dodds, 2007) and it is found that large cascades of influence are driven by a critical mass of easily influenced individuals rather than influentials.

Second, previous studies do not consider that individuals have own different threshold to adopt new ideas or products. According to Granovetter (1978) and Valente (1996), individuals vary in their willingness to take risks in adopting innovations. It was defined as threshold which means individuals' differences exist in making decision before a given actor does so (Granovetter, 1978). Individual is also classified into four categories by one's own threshold (Valente, 1996): early adopters, early majority, late majority, and laggards. Despite individual's threshold, pertinent literatures treat all the blog social ties among bloggers and readers equally. In this sense, they have not reflected the importance of aspects in influence studies. We address these limitations by considering the quality of blog social ties according to influenced individuals' thresholds.

3 Quantifying influence model

To quantify the influence score of blog posts, the QIM is proposed. Before modelling, we introduce some terms used in the blogosphere and describe observable engagement. From this, we define blog social ties as indicators of QIM and the concept of QIM is shown along with a method to compute it.

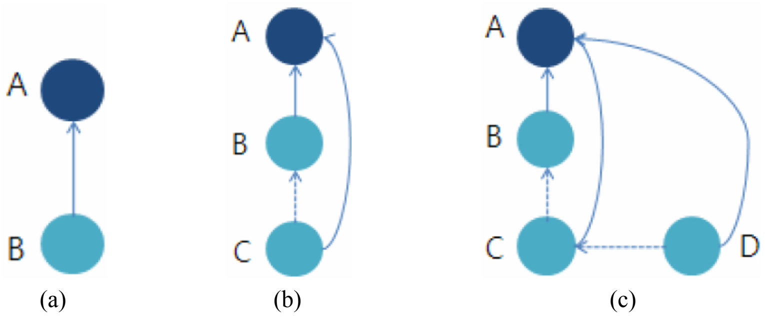
3.1 General terms used

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 behaviour. Thus, it needs to define the terms of engagements used in this paper.

1. *Comment*: Comments are usually texts published by readers in the 'comments space'. They have a link for the reader's own blog.
2. *Sympathy*: Expressing sympathy appears in various types according to blog site, e.g., in Xanga[2], 'recommendation' is used, in Digg[3], 'voting' is used so that users can give their votes in the form of a dig, and in Naver blog, '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 behaviours. We refer this as 'sympathy'.
3. *Scrap*: When readers want to take blog posts, they engage by 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 – Figure 1(a) node B creates a link to node A by scrapping. Figure 1(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 Figure 1(c), this process is also conducted in the case of 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'.

Figure 1 Cascades of blog post via scrapping (see online version for colours)



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 s 'scrap'.

3.2 Issues of popular blogger

Blog domains list 'A-List' bloggers based on traffic information, such as the number of comments, recommendations, page views and so on. However, these criteria are merely simple summations of observable statistical criteria, and do not reflect the quality of the blogger's social ties. In the case of the Naver blog, 'topic-centred popular blogs' are listed for a specific topic according to the following criteria:

1 How long the blogger has been operating his/her blog?

2 How many blog posts the blogger has written and how frequently (s)he posts new blogs?

3 How many comments and sympathy his/her blog posts receive, and how many readers view and scrap the blog posts?

4 How many online neighbours the blogger has?

Based on these criteria, 'A-List' bloggers are listed for every given topic. However, such information can be misleading since a post from a 'popular blogger' on any topic may become the top-rated blog for that topic, even if the blog has little to do with the topic (Arrington, 2006). There are several reasons why these statistics cannot be employed to select an influential blogger. First, bloggers who have posted sporadically over a long period of time will not command continuous attention from readers. Second, active bloggers are not necessarily influential, and influential bloggers may not always be active (Agarwal et al., 2008). Third, in gauging influence, the degree of the reader's interest is more important than the simple number of comments and sympathy. Furthermore, many comments and/or sympathy now comes from spam blogs or non-targeted readers, which certainly does not reflect a blog's popularity or influence. Finally, targeted key online neighbours for a specific topic are more influenced by a blogger's post, rather than readers who are not interested in the topic.

In sum, current criteria treat every reader equally despite each reader's varying inclination to be influenced by blogger's opinions. In the following section, we demonstrate how to weight blog social ties.

3.3 Indicators of QIM

In this section, we first define indicators of QIM. This is because large cascades of influences are driven not by influential, but by a critical mass of easily influenced individuals (Watts and Dodds, 2007), we need to define readers' behaviour when they are influenced by bloggers. 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.

3.3.1 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 (Marlow, 2004). As readers interact with bloggers by contributing in the form of a response to specific blog posts (Ali-Hasan and Adamic, 2007), the comment serves as 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 these engagements are channels of communication among bloggers and readers, they are observed as indicators of influence driven by readers. Finally, trackback is an automatic form of communication that occurs when one weblog references another (Marlow, 2004). The trackback system gives bloggers and readers an awareness of who is 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*.

3.3.2 Engagement at the system level

By scraping blog posts, readers send them to their own blogs or to another platform. 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 3-(A)-(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.

3.4 Concept of QIM

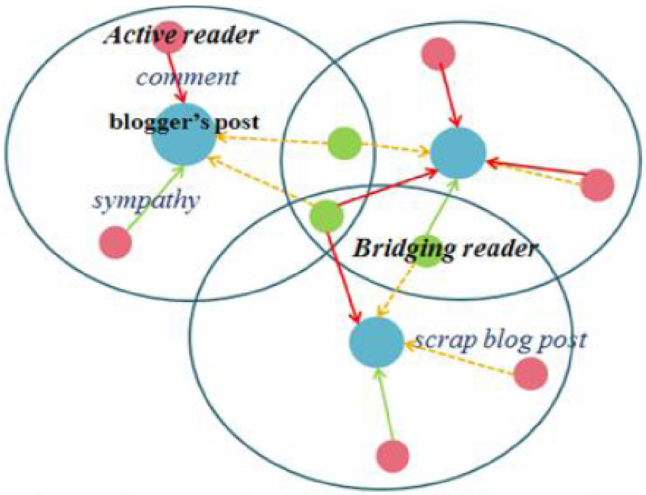
This section presents the concept of QIM, as illustrated, to measure the influence of bloggers (Figure 2), and a method to weight readers' thresholds with respect to their engagements. As described 3-(c), influenced readers take action on two levels; thus, we define an influential blogger as one having 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 i*nterpersonal 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

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.

Figure 2 Concept of QIM (see online version for colours)



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

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):

$$\text{Influence (blog post, } \Delta t) = (\text{Interpersonal similarity, Degree of information propagation}) \quad (1)$$

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.

3.4.1 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 behaviour exhibited by those they interact with (McPherson et al., 2001). This phenomenon also appears in the blogosphere. In the blogosphere, bloggers tend to build relationships with other bloggers who share similar interests and tend to visit the same blogs frequently (Herring et al., 2005). 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 (Brown et al., 2007); 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 (Brooks and Montanez, 2005; Hayes and Avesani, 2007; Li et al., 2008). 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 Jaccard coefficient (Sternitzke and Bergmann, 2009) to measure the level of similarity between a blogger and a reader (2).

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

This coefficient is calculated as the number of tags contained in a set of B's tags and R's tags and is normalised 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 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$_k$.

$$\begin{aligned}\text{Interpersonal similarity }\left(\text{blog post}_\text{k}\right) = {} & \sum \text{Jaccard}\left(\text{B, R}_\text{comment}\right) \\ & + \sum \text{Jaccard}\left(\text{B, R}_\text{sympathy}\right)\end{aligned} \tag{3}$$

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

3.4.2 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, neighbours in some network (Granovetter, 1978). 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 3-(A)-(3). More precisely, we describe this process as follow:

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 scrapped a blog post$_1$, J_k denotes the number of times blog post$_k$ was scrapped by that reader, and WJ 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 case, the degree of information propagation of blog post$_k$ is calculated as follows (4) (see Table 1).

$$\text{Degree of information propagation }\left(\text{blog post}_k\right) = \frac{W_I + W_J + W_K}{\sum_{i=1}^{N} W_i} \tag{4}$$

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

$$\text{Degree of information propagation }\left(\text{blog post}_k\right) = \frac{\sum \text{Weighted value of reader who scrapped blog post}_k}{\sum \text{Bridging reader's weighted value}} \tag{5}$$

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

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							$\sum_{i=1}^{N} W_i$

3.5 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 topic. The method of setting the weight is described in 4-(B). The completed formula of the QIM is expressed below (6) and Γ_1 and Γ_2 denote the weights.

$$\begin{aligned}
&\text{Influence }\left(\text{blog post}_k, \Delta t\right)\\
&= \Gamma_1 \cdot \text{Interpersonal similarity}\left(\text{blog post}_k\right)\\
&+\Gamma_2 \cdot \text{Degree of information propagation}\left(\text{blog post}_k\right)\\
&= \Gamma_1 \cdot \left\{\sum \text{Jaccard}\left(\text{B}, \text{R}_{\text{comment}}\right) + \sum \text{Jaccard}\left(\text{B}, \text{R}_{\text{sympathy}}\right)\right\}\\
&+\Gamma_2 \cdot \frac{\sum \text{Weighted value of Reader who scrap blog post}_k}{\sum \text{Weighted value of bridging reader}}
\end{aligned} \tag{6}$$

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 the 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, {iscore(blog post$_1$), iscore(blog post$_2$), …, iscore(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 bloggeri 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).

$$\text{Influence}\left(\text{blogger}_i\right) = \text{average}\big(\text{iscore of blogger}_i\text{'s influential blog post}_k, \\ \text{where } 1 \leq k \leq M\big) \tag{7}$$

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.

4 Empirical study

In this section, we evaluate QIM. We first define a popularity model to compare popularity with influence. Then, we present how to collect data and how to fix two weights Γ_1 and Γ_2. The result of evaluation is shown and we compare the rank of influential bloggers with that of popular bloggers.

4.1 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-centred 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-centred 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.

4.2 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 of 2009, from the 10th to the 30th of September 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 (Watts and Dodds, 2007), 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, six 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 1.7137while 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

4.3 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 list; and this set of bloggers is considered to comprise those who are both influential and popular. According to our calculations, the most influential blogger on the Naver blog was only the 27th most popular blogger, and the second most influential blogger was also in the 20th rank in the popularity model. In contrast, all three of the most popular bloggers failed to make the list of most influential bloggers. These statistics indicate that popular bloggers are not necessarily influential, and influential bloggers are not always popular. More accurate investigation is described in the following section.

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 (Watts and Dodds, 2007). Thus, tracking the top 3 influential bloggers is reasonable (Table 3).

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	2,855
2nd	16	19.3391	10	2,770.214
3rd	22	11.6100	6	1,963.25

4.4 A closer look at influential/popular bloggers

4.4.1 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.

4.4.2 Qualitative indicators from the most inlfuential/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

2 whether or not bloggers have blog social ties with other influential bloggers.

Indicator (1) denotes that R_{comment} 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 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. 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 $\mathrm{R}_{\mathrm{comment}}$ per post and 0.0625 non-active $\mathrm{R}_{\mathrm{sympathy}}$ per post, whereas the top 3 popular bloggers have 12.647 non-active $\mathrm{R}_{\mathrm{comment}}$ per post and 1.647 non-active $\mathrm{R}_{\mathrm{sympathy}}$ 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 highly 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.

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 three types of bloggers

	Influential bloggers		*Influential and 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

5 Validation

As shown in a recent study (Agarwal et al., 2008), 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.

5.1 Criteria for validation

In this paper, *influencers* are defined as those who have influential power and can make others change their thinking or behaviour (see Section 1-B). Thus, we validate this by showing that influential bloggers make more readers' thinking or behaviour 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 behaviours 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

2 making link citations.

5.2 Setup 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.

5.3 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 pots. In the case of Rsympathy, 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.

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

This result presents that the top 3 influential bloggers have fewer $R_{comment}$ and $R_{sympathy}$. However, their readers refer to influential bloggers far more frequently within their blog posts and make far more link citations of an influential blogger's blog posts than do readers of top 3 popular bloggers. On the other hand, despite having a greater number of readers $R_{comment}$ and $R_{sympathy}$, popular bloggers are rarely referenced and have few link citations made by their readers. That is, the readers of popular bloggers do not function as a channel in diffusing information, but just seek information as their needs and behave as lurkers. Accordingly, we can see that simple summations of observable statistical criteria indeed cannot identify influential bloggers.

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

6 Discussion

6.1 Influential bloggers and popular bloggers

In identifying influential bloggers, the factors that make bloggers influential or popular were 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 (1973) 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 behaviours.

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 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

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.

6.2 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.

6.3 Future works

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 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. Thus, we can propose adjusting QIM in order to identify influencers on other social media.

For the future, we propose an ongoing project in which we use QIM to identify influencers on Twitter. With its huge and ever-growing number of users, Twitter[6] has established itself as one of the most notable micro-blogging service (Milstein et al., 2008). Twitter allows *twitterers*[7] to publish tweets with a limit of 140 characters or fewer, to which other twitterers can then follow, mention, or retweet. Twitter has sparked significant interests among researchers, and many studies have been conducted to find influential twitterers. But why is it so important to identify influencers on Twitter?

'WOM' is the process of conveying information directly from person to person, by which consumers can share their opinions about products or services. The concept of WOM has been greatly enhanced by the Web, where it is referred to as electronic WOM (eWOM), and one of the most important forms of eWOM marketing involves microblogs, such as Twitter. Twitter, which is the by far the most popular microblog, allows twitterers to share their thoughts with anyone in the world who has access to the internet. Because of the unprecedented scale of its communication network, Twitter deserves serious attention as a form of eWOM (Jansen et al., 2009).

Based on Twitter's vast marketing potential, there have been several attempts to measure the influence of twitterers. Perhaps the most popular metric is the total number of followers, which suggest that the more followers a twitterer has, the more influence (s)he has on Twitter. This logic emerged from the notion that a twitterer with more followers will reach more people with his/her tweets. However, this metric overlooks several factors which typically cause twitterers to follow other people. The choice to follow other twitterers does not necessarily stem from interests or popularity. Two patterns of following on Twitter have been observed (Weng et al., 2010). First, each twitterer randomly follows someone, and those being followed follow back just for the sake of 'courtesy'. Second, it might contradict the first pattern. 'Homophily' among twitterers indicates that a twitterer may follow others since they have common interests.

Because the reciprocity observed in *following* on Twitter has various meanings and contexts, the quantity of followers is not a sufficient figure for identifying influencers on Twitter. The content of twitterers' interactions must also be considered.

Another popular metric is the ratio between the numbers of followers a twitterer has versus the number of other people that the twitterer follows. But this is believed to reveal information about types of twitterers, rather than influence on Twitter (Leavitt et al., 2009). Three types of ratio have thus far been studied:

1 Infinity: A twitterer with an infinite or near-infinite ratio of followers to followees might be interpreted as someone who focuses primarily on the material aspect of Twitter.

2 1 (equal or near-equal): A twitterer with an equal or nearly equal amount of followers and followees might be considered a conversationalist.

3 Zero: A twitterer with low followers and high followees might be considered a spammer.

Based on this empirical analysis, it is apparent that simply considering the quantity of followers or followees cannot be the precise criteria for finding influencers on Twitter.

The Web Ecology Project (Leavitt et al., 2009) analyses 12 twitterers who are widely perceived to be among the more influential users on Twitter. The authors use several criteria to rank twitterers, such as the average content spread per tweet, and average conversation activity per tweet. TunkRank (Tunkelang, 2009) demonstrates a mechanism for measuring the influence of twitterers. The authors compute the influence of twitterers by determining the expected number of twitterers who will read a tweet. Thus, they focus on measuring how far a twitterer's tweet is likely to propagate. TwitterRank, an extension of the PageRank algorithm has also been proposed to measure a user's influence on

Twitter (Weng et al., 2010). Weng et al. (2010) tracked the top 1,000 Singapore-based twitterers, along with their followers and followees.[8] They argued that the reciprocity commonly found among Twitter followers can be explained by 'homophily', which is the tendency of individuals to associate and bond with similar others (Jansen et al., 2009). However, the PageRank algorithm is not applicable to the blogosphere (Hayes and Avesani, 2007), so there is still a need to consider interaction patterns specific to Twitter, rather than the Web. Another challenge is that the reciprocity appears only incomplete data on Twitter. This study was limited to 1,000 Singapore-based twitterers, so the reciprocity was identified by TwitterRank. It might not be applicable to other groups of twitterers.

As stated, previous studies have some limitations which we would address. All the papers discussed above assume the basic metrics on Twitter and propagation models. On twitter, a communication path is the tweet, so it is necessary to analyse the content of tweets and twitterers' interactions with a qualitative approach. By conducting future work on Twitter, we can observe interaction patterns on Twitter and help marketers reach the most potential twitterers through viral marketing. This study would be significantly different from our previous analysis about influence in the blogosphere (Moon and Han, 2010), so we could derive new insights on the influence and how it varies according to the online platforms.

7 Conclusions

This paper showed how to model interaction patterns among bloggers and readers by introducing the theories of sociology. By developing a new metric, QIM, we presented that details pertaining to who influences whom as well as how, and what make bloggers influential or popular. We also proposed that how to differentiate influence from popularity in the blogosphere. Although considering the same engagements of readers in both QIM and popularity model, weighting blog social ties plays a pivotal role in identifying influential bloggers. This study has important main three contributions:

1. We attempted to position this work at the intersection of social media analysis and the social science. To the best of our knowledge, it is the first report to merge 'homophily' and 'threshold' concept in the blogosphere. The role of weak ties and strong ties in the blogosphere were reflected to QIM by giving different value. Evaluating QIM shows that weak ties are required to gain in both influence and popularity. Based on weak ties, and based on weak ties, strong and homophilous ties function as the key factors for differentiating influence from popularity.

2. We proposed 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. That is, with the guidelines of sociology, we presented a new method to compute reader's threshold to be influenced with tag-similarity and degree of information propagation.

3. We showed what makes bloggers influential or popular. If a blogger just writes a wide range of hot topics without specialty, (s)he could temporarily attract high traffic but fail to gain influence on a specific genre.

Additionally, our work lays the foundation stone of a qualitative approach to finding influential bloggers. We developed QIM with an interdisciplinary approach rather than simply analysing the topology of blogosphere, and it indicates that this approach can give more productive answers in finding influencers as a whole. In this sense, our work would be also useful for sociologists to test existing theories of sociology in the blogosphere.

Acknowledgements

We thank NHN for providing us with blog data and appreciate the anonymous reviewers for reviewing and giving feedback on this paper.

References

Adar, E., Zhang, L., Adamic, L.A. and Lukose, R.M. (2004) 'Implicit structure and the dynamics of blogspace', in *Workshop on the Weblogging Ecosystem*, May, New York, NY, USA.

Agarwal, N., Liu, H., Tang, L. and Yu, P.S. (2008) 'Identifying the influential bloggers in a community', *Proceedings of the First ACM International Conference on Web Search and Data Mining*.

Ali-Hasan, N. and Adamic, L.A. (2007) 'Expressing social relationships on the blog through links and comments', *Proceedings of International Conference on Weblogs and Social Media*.

Arrington, M. (2006) 'Finally! Bloglines blog search', available at http://www.techcrunch.com/2006/05/31/askcombloglines-launch-blog-search.

Blackwell, R.D., Miniard, P. and Engel, J. (2001) *Consumer Behavior*, South-Western, Mason.

Bott, H. (1928) 'Observation of play activities in a nursery school', *Genet. Psychol. Monogr.*, Vol. 4, pp.44–88.

Brooks, C.H. and Montanez, N. (2005) 'An analysis of the effectiveness of tagging in blogs', in *2005 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs. AAAI*, March.

Brown, J., Broderick, A.J. and Lee, N. (2007) 'Word of mouth communication within online communities: conceptualizing the online social network', *Journal of Interactive Marketing*, Vol. 21, No. 3.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. and Suri, S. (2008) 'Feedback effects between similarity and social influence in online communities', *ACM KDD*.

Friedkin, N.E. (1998) *A Structural Theory of Social Influence*, Cambridge University Press.

Gladwell, M. (2000) 'The tipping point; how little things can make a big differences', *Abacus Books*, ISBN: 0349113467.

Granovetter, M. (1973) 'The strength of weak ties', *American Journal of Sociology*, May, Vol. 78, pp.1360–1380.

Granovetter, M. (1978) 'Threshold models of collective behavior', *American Journal of Sociology*, Vol. 83, pp.1420–1443.

Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. (2004) 'Information diffusion through blogspace', *International World Wide Web Conference*, pp.491–501.

Hayes, C. and Avesani, P. (2007) 'Using tags and clustering to identify topic-relevant blogs', *International AAAI Conference on Weblogs and Social Media*.

Herring, S.C., Kouper, I., Paolillo, J.C. and Scheidt, L.A. (2005) 'Conversations in the blogosphere: an analysis 'from the bottom up'', *Proceedings of the Thirty-Eighth Hawaii's International Conference on System Sciences (HICSS-38)*.

Huston, T.L. and Levinger, G. (1978) 'Interpersonal attraction and relationships', *Annual Review of Psychol.*, Vol. 29, pp.115–156.

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A. (2009) 'Twitter power: tweets as electronic word of mouth', *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 11, pp.2169–2188.

Katz, E. and Lazarsfeld, P. (1955) *Personal Influence: The Part Played by People in the Flow of Mass Communications*, The Free Press, New York.

Kempe, D., Kleinberg, J.M. and Tardos, E. (2003) 'Maximizing the spread of influence through a social network', in *KDD*, pp.137–146.

Kempe, D., Kleinberg, J.M. and Tardos, E. (2005) 'Influential nodes in a diffusion model for social networks', in *ICALP*, pp.1127–1138.

Kleinberg, J. (1998) 'Authoritative sources in a hyperlinked environment', in *9th ACM-SIAM Symposium on Discrete Algorithms*.

Kritikopoulos, A., Sideri, M. and Varlamis, I. (2006) '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*, p.8.

Leavitt, A., Burchard, E., Fisher, D. and Gilbert, S. (2009) 'The influentials: new approaches for analyzing influence on Twitter', Web Ecology Project, available at http://tinyurl.com/lzjlq.

Li, X., Guo, L. and Zhao, Y.E. (2008) 'Tag-based social interest discovery', *International World Wide Web Conference*.

Marlow, C. (2004) 'Audience, structure, and authority in the weblog community', *International Communication Association Conference*, 27 May–1 June, New Orleans, LA.

McPherson, M., Smith-Lovin, L. and Cook, J. (2001) 'Birds of a feather: homophily in social networks', *Annual Review of Sociology*, Vol. 27.

Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B. and Magoulas, R. (2008) 'Twitter and micro-messaging revolution: communication, connections, and immedicacy-140 characters at a time', O'Reilly Report.

Moon, E.Y. and Han, S.K. (2010) 'A qualitative method to find influencers using similarity-based approach in the blogosphere', *Proceedings of IEEE International Conference on Social Computing*.

Page, L., Brin, S., Motowani, R. and Winograd. T. (1998) 'The Pagerank citation ranking: bringing order to the web', Technical report, Stanford Digital Library Technologies Project.

Qualman, E. (2009) *Socialnomics: How Social Media Transforms the Way we live and do Business*, John Wiley and Sons.

Scooble, R. and Israel, S. (2006) *Naked Conversations: How Blogs are changing the Way Business Talk to their Customers*.

Sternitzke, C. and Bergmann, I. (2009) 'Similarity measures for document mappin: a comparative study on the level of an individual scientist', *Scientometrics*, Vol. 78, No. 1.

Strang, D. and Soule, S. (1998) 'Diffusion in organizations and social movements', *Annual Review of Sociology*.

Tunkelang, D. (2009) *TunkRank: Measuring Influence by How Much Attention Your Followers Can Give You*, available at http://tunkrank.com.

Valente, T.W. (1996) 'Social network thresholds in the diffusion of innovations', *Social Networks*, Vol. 18, pp.69–89.

Van Alstyne, M. and Brynjolfsson, E. (2005) 'Global village or CyberBalkans: modeling and measuring integration of electronic communities', *Management Science*, Vol. 51, No. 6, pp.851–868.

Watts, D.J. and Dodds, P.S. (2007) 'Influentials, networks, and public opinion formation', *Journal of Consumer Research*, University of Chicago Press.

Wellman, B. (1929) 'The school child's choice of companions', *Journal of Educational Research*, Vol. 14, pp.126–132.

Weng, J., Lim, E-P., Jiang, J. and He, Q. (2010) 'TwitterRank: finding topic-sensitive influential Twitterers', in *ACM WSDM*.

Notes

[1] Available at http://blog.naver.com.

[2] Available at http://www.xanga.com.

[3] Available at http://www.digg.com.

[4] Available at http://www.google.com.

[5] Available at http://www.technorati.com.

[6] Available at http://www.twitter.com

[7] In this paper, we refer to Twitter users as *twitterers*.

[8] 'Followees' are those who are being followed on Twitter.