Measuring the Influence of Bloggers in Their Community Based on the H-index Family

Dinh-Luyen Bui, Tri-Thanh Nguyen*, and Quang-Thuy Ha

Vietnam National University, Hanoi (VNU),
University of Engineering and Technology (UET)
{luyenbd_54,ntthanh,thuyhq}@vnu.edu.vn

Abstract. Nowadays, people in social networks can have impact on the actual society, e.g. a post on a person's space can lead to real actions of other people in many areas of life. This is called social influence and the task of evaluating the influence is called social influence analysis which can be exploited in many fields, such as typical marketing (object oriented advertising), recommender systems, social network analysis, event detection, expert finding, link prediction, ranking, etc. The h-index, proposed by Hirsch in 2005, is now a widely used index for measuring both the productivity and impact of the published work of a scientist or scholar. This paper proposes to use h-index to measure the blogger influence in a social community. We also propose to enhance information for h-index (as well as its variants) calculation, and our experimental results are very promising.

Keywords: social network, influence of blogger, h-index.

1 Introduction

In real life, people usually tend to consult others (e.g. family members, relatives, friends, or experts) before making decisions, especially important ones. As reviewed by [1], 83% of people ask others for experience before trying a restaurant, 71% of people do the same before buying a prescription drug or visiting a place, and 61% of people talk to others before watching a movie. Thanks to the characteristic of social networks that makes the information distribution almost at real-time, it leads to the change of daily behaviors of people who participate in a social network. For example, before buying a certain product (e.g. a mobile phone), people tend to search for others' available comments, experiences or evaluation on the product. As a result, if the content of a user's post is interesting and reliable, it can have a certain impact on other people in that network community. In other words, people have one more source of consultant affecting their daily habits.

A recent typical example that shows the influence of a user on a social network on economy is two tweets of Carl Icahn on Tweeter in August 2013: "*We currently have a large position in APPLE. We believe the company to be extremely undervalued.*

* Corresponding author.

T.V. Do et al. (eds.), *Advanced Computational Methods for Knowledge Engineering*,
Advances in Intelligent Systems and Computing 282,
DOI: 10.1007/978-3-319-06569-4_23,

Spoke to Tim Cook today. More to come", and "*Had a nice conversation with Tim Cook today. Discussed my opinion that a larger buyback should be done now. We plan to speak again shortly.*" The two tweets had a big impact on Apple's stock market. The value of Apple's stocks increased more 12 billion US dollars with about 200.000 stock transactions soon after the appearance of the tweets. Such fact raised a new topic called *social influence analysis* which evaluates the influence capacity of a user (in a social network) on the others. In other words, it evaluates how much an action (described in a user's post) can lead to certain actions of other people in the community as well as real society.

N. Agarwal et al. [1] proposed a model (called iFinder) which attempts to figure out top k influential bloggers having highest scores. The key idea is to score all the posts of bloggers in a community, and select the highest score of one's posts to be his/her influence score (more details of the model will be given in Section 2). Naturally, influence score should be a value that is accumulatively calculated and increased over new posts. Hence, if the influence score relies on only one post, we do not take the contribution of other posts into account, and it does not seem reasonable. In addition, such a score is not reliable in some situations, such as spamming in which spammers simply make some effort to increase the score of only one of his posts. Though the authors claimed that it is possible to use the mean score of all posts as the influence score, this calculation method, again, has a drawback, i.e. it takes into account both influential and non-influential posts. Finally, based on the fact that the life time (time to have attention) of posts in social networks is short, if we rely on a single post score, and when this post is obsolete, it is not reasonable to use its score as blogger's score.

In this paper, we propose to apply the h-index [8] to calculate the influence score of bloggers which will better reflect the reality. The h-index was proposed by Hirst to measure both the productivity and impact of the published papers of a researcher. If a researcher has N published papers in which there are h papers ($h \leq N$) each of which has at least h (inbound) citations, then his h-index is h. It is easily noted that the productivity is the number of papers (h) that have impact (as the number of citations h). When the h-index is applied to rank bloggers, we do not rely on a single post anymore, and also calculate non-influential (or less influential) posts.

However, as we can see, the h-index does not take outbound citations into account. This is not appropriate for social networks where inbound and outbound links and other related information play the role of essential constructs for information navigation and distribution. In this paper, we propose to utilize the post score of iFinder which incorporates several properties (besides inbound links) in the first step of h-index calculation.

The next problem we faced in this work is that the posting score of iFinder is a real number (in the range of [0..1)) which cannot be directly used for the h-index calculation. We use two methods to convert a real number post score to an integer for h-index calculation. Finally, since the h-index was introduced, there have been several proposed variants with improvements. In this work, we also calculated influence score using h-index variants for evaluation.

The rest of the paper is organized as follows. Section 2 briefly introduces related work. Section 3 presents our model to calculate influence scores. Section 4 shows the experimental results and evaluation. Finally, Section 5 concludes the paper and gives some potential future directions.

2 Related Work

2.1 Influential Blogger Identification

The people whose experiences, opinions, and suggestions are sought after are called the influentials [2]. As stated by M. Momma et al. [13], social influence has two forms: the first one is the action (or behavior) (stated in the post) itself, and the second is that this action can lead to the action of other people. The second form is the object of this paper that reflects the impact of influential on other individuals in the community. As reviewed in [1], the identification of the influential bloggers can benefit all in developing innovative business opportunities, forging political agendas, discussing social and societal issues, and lead to many interesting applications [5, 7, 10, 11, 12, 14]. For example, the influentials are often market-movers. Since they can influence buying decisions of the fellow bloggers, identifying them can help companies better understand the key concerns and new trends about products interesting to them, and smartly aspect them with additional information and consultation to turn them into unofficial spokesmen. Approximately 64% advertising companies have acknowledged this phenomenon and are shifting their focus toward blog advertising. As representatives of communities, the influentials could also sway opinions in political campaigns, elections, and aspect reactions to government policies. Tracking the influentials can help understand the changing interests, foresee potential pitfalls and likely gains, and adapt plans timely and pro-actively (not just reactively). The influentials can also help in customer support and troubleshooting since their solutions are trustworthy in the sense of their authority in term of being influentials.

The influential blogger identification can be roughly defined as: *Given a set of M bloggers* (*in a certain community*)*, find out K* ($K \leq M$) *bloggers who have highest scores* (*according to a certain estimation*).

Nitin Agarwal et al. [1] proposed a model called iFinder for calculating blogger influence score, which will be introduced in detail in Section 3.

2.2 H-index Family

In this section, we briefly introduce h-index as well as its variants which will be used in our research.

The h-index was proposed by Hirsch in 2005 [8] to be used as an index of a scientist or scholar. It is defined as follows:

A scientist has index h if h of his/her N_p *papers have at least h citations each, and the other* (N_p *h) papers have no more than h citations each.*

Let C be the set of top most cited papers of a scientist, U be the set of all the scientist's papers, $cite(p)$ be the function returning the number of citations to paper p, then the h-index h of the scientist is defined as follows:

$$h = arg \max_{C \subseteq U} |C|$$

$$such\ that\ \forall p \in C, cite(p) \geq |C| \wedge\ \forall p \in U \backslash C, cite(p) < |C| \quad (1)$$

For example, a scientist published 6 papers. Assuming that for two top most cited papers, each has 6 references, while each of the rest has 2 references. Then the h-index of this scientist is 2. The common sense of the h-index is that it increases as the number of papers and citations accumulate, and thus it depends on the 'academic age' of the scientist. It also has quantitative aspect: As reviewed by the author, for physicists, a value for h of about 12 might be typical for advancement to associate professor at major research universities. A value of about 18 could have a full professorship; 15–20 could gain a fellowship in the American Physical Society; and 45 or higher could mean membership in the United States National Academy of Sciences. This indicates the h-index to be a stable and consistent estimator of scientific achievement. Thus, it is currently used to rank objects bigger than a person, such as a department, a university, a country or a journal.

L. Egghe [6], in 2006, argued that h-index has a problem of assigning the same weight to all papers that contribute to h-index, since when a researcher has the index h, and one of his papers has much more citations than h, this paper contributes the same weight as that of the top h papers. Egghe proposed another index called *g-index* as follows:

Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number g such that the top g articles received a total of at least g^2 citations.

Let C be the set of g top most cited papers ($C = \{p_1, p_2, \dots, p_{|C|}\}$) the formula for *g-index* can be defined as follows:

$$g\text{-}index = arg \max_{C \subseteq U} g\ \ such\ that\ g^2 \leq \sum_{i=1}^{|C|} cite(p_i) \quad (2)$$

We can notice that total number of top g papers is used in *g-index* calculation, hence, a paper of higher number of citations contributes more weight to the index than a smaller one. With the same argument as that of Egghe, Jin [9], in 2006, proposed another variant of *h-index* called *A-index*. If a researcher has the *h-index* h constructed from the set C of top most cited papers ($C = \{p_1, p_2, \dots, p_h\}$), then *A-index* is defined as follows:

$$A\text{-}index = \frac{1}{h} \sum_{j=1}^{h} cite(p_j) \quad (3)$$

However, this formula still has a problem as stated in [4]. Consider the following situation: an author X_1 published 20 papers, in which one paper has 10 citations while each of the rest has only one citation; another author X_2 published 30 papers, in which one paper has 30 citations while each of the rest has 2 citations. Naturally, author X_2

should be considered to be better than X_1. Nonetheless, H-indices of X_1 and X_2 are 1 and 2, correspondingly, whereas, the *A-indices* of the two authors X_1 and X_2 are 10 and 6, correspondingly. This drawback comes from the fact that *A-index* formula has a division by h. Suppose an author has h-index h, based on the set of h top most cited papers, J. BiHui et al. [4], in 2007, proposed another one called R-index which is defined as,

$$R\text{-}index = \sqrt{\sum_{j=1}^{h} cite(p_j)} \tag{4}$$

Peter Vinkler [15], in 2009, proposed $\pi\text{-}index$ to improve the h-index. Suppose the total number of papers of a scientist is T that are sorted in the acceding order of number of citations, let the elite set P_π be $\lfloor\sqrt{T}\rfloor$ top most cited papers, $C(P_\pi) = \sum_{p\in P_\pi} cite(p)$, then $\pi\text{-}index$ is defined as follows:

$$\pi\text{-}index = 0.01C(P_\pi) \tag{5}$$

Due to the limitation of *A-index*, we will not use it in our experiments. The h-index is used to measure the productivity as well as impact in the whole academic life of a scientist, so it should increase over time. However, when it is used to rank bloggers, we can calculate h-index of a blogger based on the data in a certain duration (not the whole), so that it can increase or decrease depending on the data. In other words, it is possible to compare the influence of the blogger in different time durations.

3 Using the H-index to Measure Influence

3.1 Rationale

Based on the intuition that when paper *A* refers to another one *B*, *A* tends to borrow information from *B*. In other words, *B* is an information source. The more references *B* has, the more interesting it is. Thus, the h-index bases only on inbound citation information for calculation. The situation is completely changed in World Wide Web or social networks. Let's analyze some important properties other than inbound reference (citation) which should be considered in index calculation: a) *Outbound links* also play important roles in information navigation or distribution. For a website of an organization, the home page has a crucial role, because it stores the links as a map to guide users to navigate to their expected pages. For social network sites, such as Twitter, when a user *A* follows (or links) to another one *B*, then *B*'s new tweets will appear in (or be distributed to) *A*'s home page. In this case, the outbound link (from *A* to *B*) servers as a clue for information distribution. b) *The content of the post* (or webpage or tweet) is an important property in the context whether it is a hot/contemporary topic in the real world. This may be the most important aspect, however, it is the most difficult aspect to estimate. c) *Response*: a post can attack others to respond in a form of comments/discussions. The more comments a post has, the more interesting it tends to be. d) *Related information* of the user in real life (e.g. the position of job or

expertise): as seen in the example of Icahn's tweets, the position of Icahn has a big effect on the others. However, this information is difficult (even impossible) to obtain. e) *The number of reads* (or visits): may indicate a certain interesting level of the post. f) *Activeness*: an active user may usually have new information to post.

From this discussion, we propose to integrate some more properties (information) into h-index calculation. After a review, we noticed that iFinder has exploited and incorporated some additional properties in their model, thus, we reuse the calculation model of iFinder as the first step for h-index calculation. Before introducing our model, we briefly present the iFinder model in the next subsection.

3.2 iFinder Model

Influential Blogger definition: A blogger is influential if s/he has at least one *influential blog post*

For a blogger b_k who has N blog posts $\{p_1, p_2, ...,p_N\}$; denote the influence score of i^{th} post as $I(p_i)$, then b_k influence index (*iIndex*) is defined as follows:

$$iIndex = \arg\max_{i=1..N} I(p_i) \tag{6}$$

A blog post p_i is deemed influential iff $I(pi) \geq \alpha$, where α is a threshold determined at the calculation time based on the number of the most influential bloggers.

Problem Statement: Given a set U of M bloggers $\{b_1, b_2, \dots, b_M\}$, the problem of identifying influential bloggers is defined as determining an ordered subset V of K [1] most influential bloggers (with highest *iIndex* values): $V = \{b_{j1}, b_{j2}, \dots, b_{jK}\}$ sorted by their *iIndex* in the descending order such that $V \subseteq U$ and $K \leq M$, i.e. $iIndex(b_{j1}) \geq iIndex(b_{j2}) \geq \dots \geq iIndex(b_{jK})$. In this problem, we can see that the threshold α is equal to $iIndex(b_{jK})$.

As stated by K. Apostolos et al. [3], the graphs (based on the links) of blog sites are very sparse, hence, it is not suitable to rank blog posts using Web ranking algorithms (e.g. the PageRank algorithm). N. Agarwal et al. [1] proposed an alternative model to identify influential bloggers called iFinder which is described below.

The initial properties (or parameters) used to calculate the influence score of a blog post are: its set of inbound links (ι); its set of comments (γ); its set of outbound links (θ); and the length of the post (λ).

Let I(p) denote the influence score of a node p (e.g. a blog post) in the graph representing a blog site, then the *InfluenceFlow*(.) across that node is given as follows:

$$InfluenceFlow(p) = w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n) \tag{7}$$

where w_{in} and w_{out} are weights used to adjust the contribution of inbound and outbound influence, respectively; p_m $(1 \leq m \leq \theta)$ is a post that has a link to p; p_n $(1 \leq n \leq \theta)$ is a post that is referred by p;

[1] K is a user specified parameter.

InfluenceFlow(.) measures the difference between the total incoming influence of all inbound links and the total outgoing influence by all outbound links of the blog post p. It accounts for the part of influence of a blog post that depends upon inbound and outbound links. The intuitive aspect of this function is that: if a blog post is referred by another one, then it seems to have novelty, and then it gets bonus score; however, when a post links to another post, then its content seems to '*borrow*' information from an external source, and it gets penalty score.

In addition, the post's comments also indicate that the post is interesting or has novelty, hence influence $I(p)$ is proportional to the number of comments (γ_p),

$$I(p) = w_c\gamma_p + InfluenceFlow(p) \tag{8}$$

where w_c is the contribution weight of the total number of comments γ_p on the post p.

The last parameter is the length of the post λ_p. It is not simply to use λ_p as a weight, Agarwal proposed to convert λ_p to a weight by a function $w(.)$, and the final formula for $I(p)$ (from Eq. 8) is written as follows:

$$I(p) = w(\lambda_p) \times \left(w_c\gamma_p + InfluenceFlow(p)\right) \tag{9}$$

The influence score of each post $I(p)$ is normalized in the range of [0..1).

Given a set U of M bloggers who have a set P of N blog posts $P = \{p_1, p_2, \dots, p_N\}$, denote A as the adjacency matrix, where each entry A_{ij} represents the link between the post p_i and p_j. i.e. if p_i refers to p_j, then A_{ij}=1; otherwise A_{ij}=0. Matrix A represents the outbound links among posts, consequently, A^T represents the inbound links among the posts. Define the vectors of post length $\vec{\lambda}$, comments $\vec{\gamma}$, influence $\vec{\imath}$, and influence flow $\vec{f}$ as follows:

$$\vec{\lambda} = \left(w(\lambda_{p_1}), w(\lambda_{p_2}), \dots, w(\lambda_{p_N})\right)^T,$$
$$\vec{\gamma} = \left(\gamma_{p_1}, \gamma_{p_2}, \dots, \gamma_{p_N}\right)^T,$$
$$\vec{\imath} = \left(I(p_1), I(p_2), \dots, I(p_N)\right)^T,$$
$$\vec{f} = (f(p_1), f(p_2), \dots, f(p_N))^T$$

Now, Eq. 7 can be rewritten as follows:

$$\vec{f} = w_{in}A^T\vec{\imath} - w_{out}A\vec{\imath} = (w_{in}A^T - w_{out}A)\vec{\imath} \tag{10}$$

and Eq. 9 can be rewritten as follows:

$$\vec{\imath} = \mathrm{diag}(\vec{\lambda})(w_c\vec{\gamma} + \vec{f}) \tag{11}$$

Combine Eq. 10 and Eq. 11, we have

$$\vec{\imath} = diag(\vec{\lambda})(w_c\vec{\gamma} + (w_{in}A^T - w_{out}A)\vec{\imath}) \tag{12}$$

It is possible to solve the iterative Eq. 12 using *power iteration method* as described in Algorithm 1 [1].

Input: A set P of blog posts, the termination parameters: number of iteration $iter$, the similarity threshold τ
Output: The influence vector $\vec{\imath}$ representing the influence score of all the blog posts in P

Compute the adjacency matrix A
Compute vectors post length $\vec{\lambda}$, comments $\vec{\gamma}$
Initialize $\vec{\imath} = \overrightarrow{\imath_0}$
repeat
  $\overrightarrow{\imath'} = diag(\vec{\lambda})(w_c\vec{\gamma} + (w_{in}A^T - w_{out}A)\vec{\imath})$
  $iter \leftarrow iter - 1$
until $(cosine\_similarity(\vec{\imath}, \overrightarrow{\imath'}) < \tau)$ or $(iter \leq 0)$

Algorithm 1. Influence calculation (blog posts' score calculation)

After experiments, the author found out the contribution order of the 4 properties used in the iFinder model is: inbound links > comments > outbound links > blog post length, and the combination of the four gives the highest performance indicating that the selection of the four properties is suitable.

3.3 Our Model

In this section, we describe the details of our model for finding top K influential bloggers based on the h-index family. In comparison with scientific articles, the life time of posts (from the time the post appeared to the last time it was referred) in social networks is shorter, thus using the h-index family for measuring the influence is a more meaningful than the measuring method of iFinder which only bases on a single post. Since when the post represented for a blogger's influence score is obsolete, it should not be the representative anymore. Our model to identify influential bloggers is based on the h-index family, which is different from that of iFinder, we redefine an influential blogger as,

A blogger has the influence score of h if h is his/her h-index (*or its variant*) *value*.

And the influential blogger identification problem is defined as follows:

Input: A set U of M bloggers who have N blog posts and a parameter K ($K \leq M$)
Output: The set V of K top h-index bloggers.

Our model is described in Fig. 1, which has following steps:

Preprocessing: for each post, we parse each post to extract essential information for next steps, e.g. the post title; the content of the post; the length of the post; the number of inbound links; the number of outbound links; the author (blogger) of the post; the number of comments; the tags of the post; the timestamp (post time).

Post score estimation: as discussed in Section Rationale, we would like to integrate some more properties (besides inbound links). However, due to some limitation (e.g. the availability of data), we finally selected same four properties as those of iFinder, i.e., the number of inbound links; the number of outbound links; the number of comments; the post's content (estimated as the post length). We apply iFinder model to estimate the score of each post. The results of this step are the scores of each post in the range of [0..1).

Post score conversion: since the post score (returned by the previous step) in the range of [0..1) is not compatible for h-index calculation, we propose to use *binning* for transforming a post score into an integer. There are two binning methods:

- *Equal-frequency* (*or equal-depth*) *binning*: given *m* posts, equal-frequency binning method divides them into *n* bins, so that the bins have an equal number of posts. Formally, let *pos*(*p*) denote the position of post *p* in the sorted list by score in the ascending order, the bin number of *p* is $bin(p) = \lfloor pos(p)/n \rfloor$.
- *Equal-width binning*: in this method, each bin will have the same interval range of value instead of number of posts. Denote *l*, *r* as the lower and upper bounds of the target integer range, correspondingly. The interval range (*irange*) of each bin is $irange = \frac{r-l}{n}$, and the range of i^{th} bin is $[l + (i-1) * irange, l + i * irange)$ where $(1 \leq i \leq n)$. Given a post *p* then $bin(p) = i$ if $score(p) \in bin_i$

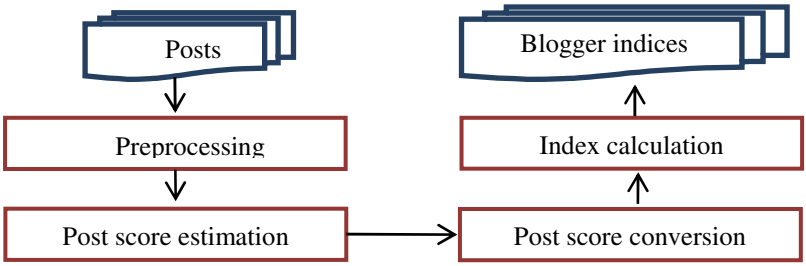

Fig. 1. The ranking model based on the h-index family

Index calculation: for each blogger, we collect the *bin*(.) values of all his posts to use as the number of citation (i.e., *cite*(.) function), and then calculate the values of the variant of h-index. After this step, we have the index of all bloggers, hence, we can sort the blogger list by their index and return *K* top highest index bloggers.

In the real world, the influence of a blogger may increase or decrease (not always increase as h-index for a scientist). However, as discussed in Section 2, it is possible to apply our model to calculate the index of a blogger based on the data subset collected at a certain duration in order to track the influence change of the blogger over time to reflect the real situation.

4 Experiments and Evaluation

4.1 Data Set and Experimental Setup

Thanks to the support of Nitin Agarwal et al. [1], we had the data set "The Unofficial Apple Weblog" (TUAW) which consists of about 10,000 blog posts from 35 bloggers. The dataset was manually investigated to rank bloggers based on their activeness. The parameter settings used in iFinder model (cf. Algorithm 1) are those recommended by the author. In equal-depth binning, we set the number of bins to 100. In equal-width binning, we set $l = 1; r = 1000; n = 1000$ (or $irange = 1$).

4.2 Experimental Results and Evaluation

We ran our model with two binning methods (i.e., equal-depth and equal-width) which both gave the same set of top 5 of most influential bloggers. To evaluate our model, similar to iFinder, we compare top 5 bloggers returned by our model (with the rank of equal-depth binning) with those of iFinder and TUAW as shown in Table 1.

Table 1. Comparison of top 5 bloggers

TUAW	iFinder	Our model
Erica Sadun	*Erica Sadun*	*Scott McNulty*
Scott McNulty	Dan Lurie	C. K. Sample, III
Mat Lu	*David Chartier*	Dave Caolo
David Chartier	*Scott McNulty*	*David Chartier*
Micheal Rose	**Laurie A. Duncan**	**Laurie A. Duncan**

As claimed by Nitin Agarwal, an influential blogger can be, but not necessarily, an active one. Thus the results returned by iFinder are not the same as top 5 active bloggers. Refer to Table 1, iFinder shares three bloggers (in italic) with TUAW, while our model shares two bloggers with TUAW (in italic), and shares 3 bloggers with iFinder. As reviewed by Agarwal, *Dan Lurie* is not active (i.e. not in the top of TUAW) but influential. Because, *Dan* has 4 influential posts and, especially, one of them writing about IPhone attacked a large number of discussion, and iFinder selects this highest post score as the influence score of a blogger resulting in *Dan* appearing in top 5. However, recalling the discussion in Section 1 that this score selection is a drawback of iFinder where spammers simply try to boost one of his posts to have a high score leading them to be influentials.

Our model did not put *Dan Lurie* in the top 5 influentials thanks to the difference in blogger score calculation. Another example is *Erica Sadun* who is marked as the first ranked influential blogger by both TAUW and iFinder. His most influential post is a keynote speech of Apple Inc. CEO Steve Jobs, which fostered a big number of comments and inbound links (two of the most influential properties contributing to the post score) giving him the highest score in iFinder model. Nonetheless, the h-index family does not rely on a single post, and assigns *Erica Sadun* a lower score in

comparison with the fifth blogger *Laurie A. Duncan*. That is also the reason why two bloggers: *C. K. Sample* and *C. K. III Dave Caolo* appear in top 5 of our model.

Observation from equal-width and equal-depth binning experiments, the two methods produced the same top 5 influential set with 4 different indexes (i.e. *h-index*, *g-index*, *r-index* and π-*index*), however, the blogger's index values are different. There are 3 different bloggers in top 10 set between the two methods indicating that top influential bloggers seem to be stable in two binning methods and 4 indexes. In addition, equal-depth binning gave higher index values than equal-width binning, though the scale of equal-width binning (in the range of [1..1000]) is larger than that of equal-depth binning (in the range of [1..100]). This is from the fact that the post scores do not distribute equally in the range but group in discrete clusters. At the moment, we haven't found out a suitable method to evaluate which index among the four is the best. This is a potential problem for our future.

Table 2. g-index of top 5 bloggers over time

Blogger	2004	2005	2006	2007
Scott Mcnulty	0	92	98	98
C. K. Sample, III	0	94	95	95
Dave Caolo	0	90	95	85
David Chartier	0	86	96	96
Laurie A. Duncan	43	87	94	94

We also carried out experiments to observe the change of blogger's influence score over time. As discussed in Section 3, we calculated the index (e.g. *g-index*) of a blogger based on a data subset (e.g. in one year duration). From the four year results of top 5 bloggers' *g-index* in Table 2, we can notice that the index can increase or decrease depending on the actual data. This means it is possible to use an index to follow the influential change of a blogger.

5 Conclusion and Future Work

In this paper, we proposed to use the h-index family for ranking bloggers in order to find out the top most influential ones. For enhancing the information used in h-index calculation, we proposed to integrate some more properties (in addition to inbound reference). The experimental results proved our proposed model are comparable to the iFinder model. Moreover, our model may avoid the drawback of iFinder model, i.e. vulnerable to spam. For the future work, we plan to integrate some more properties as discussed in Section 3, and apply our model to other domain than blogosphere, such as Facebook or Twitter.

Since the life time (the time of having attention) of a post is much shorter than that of a scientific paper, we plan to incorporate some information (e.g. the post time) in score estimation.

Another future direction is h-index threshold determination, as estimated by Hirsch in 2005 [8], a certain h-index value a physicist has can be appropriate for a certain

academic position or award (e.g. associate/full professor, cf. Section 2). We plan to figure out the threshold to judge a blogger to be influential instead of simply returning the top ranked ones.

The final future stuff is to judge which index (in the h-index family) is the most suitable for measuring influences.

Acknowledgments. This work was partially supported by the VNU Scientist links and Grant No. BB-2012-B42-29.

References

1. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Modeling blogger influence in a community. Social Netw. Analys. Mining 2(2), 139–162 (2012)
2. Akritidis, L., Katsaros, D., Bozanis, P.: Identifying the Productive and Influential Bloggers in a Community. IEEE Transactions on Systems, Man, and Cybernetics, Part C 41(5), 759–764 (2011)
3. Kritikopoulos, A., Sideri, M., Varlamis, I.: BLOGRANK: Ranking We-blogs based on Connectivity and Similarity Features. CoRR abs/0903.4035 (2009)
4. BiHui, J., LiMing, L., Rousseau, R., Egghe, L.: The R- and AR-indices: complementing the h-index. Chinese Science Bulletin 52(6), 855–963 (2007)
5. Egghe, L.: The Hirsch index and related impact measures. In: ARIST, pp. 65–114 (2010)
6. Egghe, L.: Theory and practise of the g-index. Scientometrics, 131–152 (2006)
7. Goyal, A.: Social Influence and its Applications: An algorithmic and data mining study. PhD Thesis, The University of British Columbia, Vancouver (2013)
8. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proc. of the National Academy of Sciences of the United States of America 102(46), 16569–16572 (2005)
9. Jin, B.: H-index: an evaluation indicator proposed by scientist. Science Focus, 8–9 (2006)
10. Keller, E., Berry, J.: One American in ten tells the other nine how to vote, where to eat and, what to buy. They are The Inuentials. The Free Press (2003)
11. Lee, Y., Jung, H.-Y., Song, W., Lee, J.-H.: Mining the blogosphere for top news stories identification. In: SIGIR 2010, pp. 395–402 (2010)
12. Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J., Tseng, B.L.: Splog detection using self-similarity analysis on blog temporal dynamics. In: Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web, AIRWeb (2007)
13. Momma, M., Chi, Y., Lin, Y., Zhu, S., Yang, T.: Influence Analysis in the Blogosphere. CoRR abs/1212.5863 (2012)
14. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and Passivity in Social Media. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part III. LNCS, vol. 6913, pp. 18–33. Springer, Heidelberg (2011)
15. Vinkler, P.: The pi-index: a new indicator for assessing scientific impact. Information Science (JIS), 602–612 (2009)