Identifying Opinion Leaders in the Blogosphere

Xiaodan Song, Yun Chi, Koji Hino, Belle L. Tseng
NEC Laboratories America, 10080 N. Wolfe Road, SW3-350, Cupertino, CA 95014, USA
{xiaodan, ychi, hino, belle}@sv.nec-labs.com

ABSTRACT

Opinion leaders are those who bring in new information, ideas, and opinions, then disseminate them down to the masses, and thus influence the opinions and decisions of others by a fashion of word of mouth. Opinion leaders capture the most representative opinions in the social network, and consequently are important for understanding the massive and complex blogosphere. In this paper, we propose a novel algorithm called InfluenceRank to identify opinion leaders in the blogosphere. The InfluenceRank algorithm ranks blogs according to not only how important they are as compared to other blogs, but also how novel the information they can contribute to the network. Experimental results indicate that our proposed algorithm is effective in identifying influential opinion leaders.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – *Information Filtering*; J.4 **[Computer Applications]**: Social and Behavioral Sciences – *Economics*

General Terms: Algorithms, Experimentation

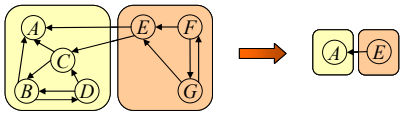
Keywords: Opinion Leader, Blog Ranking, Network Summarization

1. INTRODUCTION

Blog is a self-publishing media on the Web that allows millions of people to easily publish, read, respond, and share their ideas, experiences and expertise. The blogosphere is a fruitful media to understand people's response to events, and customers' opinions on products and services of a company, since they reflect as many topics, events, and opinions as there are people writing about them.

Comparing to traditional Web sites, the blogosphere is more conversational in style. The conversations usually start from ones who introduce new information, ideas, and opinions, then spread them down to their friends, families, and peers. *Social influence*, which describes the phenomenon by which the behavior of an individual can directly or indirectly affect the thoughts, feelings, and actions of others in a population [1, 2], is present in the conversations in the blogosphere. Those who play a crucial role in forming and reflecting the opinions of the masses are called opinion leaders in the "Diffusion of Innovation Theory" [2, 3]. Opinion leaders absorb information through various means and then transmit this information to certain opinion receivers. The important role of opinion leaders has attracted growing attention recently since massive quantities of network data are available through the Internet [4-7]. Among them, [4, 5] attempt to model influence among consumers and understand how the influence propagates in the networks. Recently, Song *et al.* studied the asymmetric inter-personal influence between opinion leaders and the masses, and proposed an asymmetric recommendation algorithm to improve the recommendation performance [6, 7]. With a great number of blogs in existence, it is valuable to discover opinion leaders to identify worthwhile readings for an individual. It would also be more effective for companies, customer services, and market researchers to seek and leverage the help of opinion leaders to understand the voices of customers, manage their brand and reputation, and promote their products.

As a motivating example, Figure 1(a) illustrates how seven blogs refer (link) to each other when they publish their opinions. Blogs *A*, *B*, *C*, and *D* discuss the same topic – *e.g. how to use Riya to find similar faces and objects on images across the web*. Then blog *E* initiates the discussion of *a rumor of Google acquiring Riya*, and links to blogs *A* and *C* that *introduce how to use Riya's visual search*. Following blog *E*, blogs *F* and *G* start to discuss this acquisition rumor. In this simple example, blog *A* and blog *E* are opinion leaders - they introduce innovative opinions and influence the opinions of other blogs. These opinion leaders capture the most representative opinions in the blog network, and thus are important for understanding and capturing the opinions in this "Riya" network (Figure 1(b)).



(a) A blog network (b) Opinion leaders

Figure 1: A motivating example

In this paper, we address the problem of how to identify the influential opinion leaders for understanding the blogosphere. Based on the characteristics of opinion leaders, we propose an InfluenceRank algorithm to rank blogs according to how important they are in the network and how novel the information they provide. The top blogs ranked by InfluenceRank tend to be more influential and informative in the network, and thus are more likely to be opinion leaders.

The rest of the paper is organized as follows. We review related work in Section 2. In Section 3, we propose our InfluenceRank algorithm. In Section 4, we demonstrate how the InfluenceRank algorithm can be used to find opinion leaders by conducting the experiments on a blog dataset. Finally, conclusions are given in Section 5.

2. RELATED WORK

In Diffusion of Innovation Theory [2, 3], two techniques have been commonly used to locate opinion leaders: the self-designation method and sociometry method [8] by designing questionnaires and interviews. Those methods are expensive and very difficult to administer and execute [8]. In contrast, the easily

CIKM'07, November 6-8, 2007, Lisboa, Portugal.

available links and content in the massive blog datasets provide the evidence of information seeking and referral behaviors and thus make the task of identifying opinion leaders possible to achieve without the necessary of designing comprehensive and costly questionnaires and interviews.

Identifying importance of nodes in a network has been an active problem in many research areas [9-11]. PageRank and HITS [9, 11] are two popular and successful link-based webpage ranking algorithms to determine the importance of web pages. As a variant of the eigenvector centrality measure, the intuition behind PageRank is that the importance of a web page in a network is proportional to the combined importance of its neighbors [9]. The intuition behind HITS is that good hubs usually point to good authorities and good authorities are usually pointed by good hubs [11]. Thus the authority score of a web page in a network is proportional to the combined hub scores of those web pages pointing to it, and its hub score is proportional to the combined authority scores of those web pages that it points to. Both PageRank and HITS provide ways of calculating the importance of nodes in a network. However, how novel the information each node contributes to the network, which is an essential property of opinion leaders, is not taken into account in both models.

Novel information detection has long been a subject of great importance in information retrieval and filtering [12]. The first story detection task [12] associated with topic detection and tracking is to detect new stories that have not been published previously. In information filtering area, Zhang *et al.* provides a set of redundancy measures to evaluate whether a relevant document contains novel information [13]. However, in these research studies, the data to be analyzed are document streams and thus the link structure is not explored in their framework. In comparison, conversations in the blogosphere via hyperlinks are important for understanding the referral behaviors and information diffusion process. Therefore, the above work cannot be directly used to identify opinion leaders.

In related work of information diffusion in the blogosphere, Kumar *et al.* identified bursty community of blogs based on community extraction and burst analysis [14]. Gruhl *et al.* analyzed temporal characteristics of blog postings on a specific topic [15]. Adar *et al.* studied the explicit and implicit link structure and the dynamics of the blogosphere [16]. Nakajima *et al.* attempts to find agitators, who stimulate discussions; and summarizers, who summarize discussions, by thread detection [17]. However, their purpose was not to acquire opinion leaders, who are key players in information diffusion process and novel information contributors in the blogosphere.

3. INFLUENCERANK

Opinion leaders are novel information contributors and also influence the opinions of other blogs. Thus, both information novelty and the importance of its position in the blogosphere are essential characteristics to a blog to determine its leadership. In this section, we discuss how to measure the information novelty among blogs, and then propose our InfluenceRank algorithm.

When an entry in a blog is generated, the content of the entry could either come from other blog posts that it links to, or be generated by its own - from other media, such as news or TV programs, or from its original ideas as novel information to the blogosphere. One can imagine there is an extra source for this blog to contribute novel information to the network. In our study, we model this extra source as a hidden node that is linked by this blog. The hidden node allows the blog to generate extra information that does not depend on its connections to others, and thus represents the capability of generating novel information, denoted as $Nov(\cdot|Out(\cdot))$, where $Out(\cdot)$ denotes the set of blogs that blog $\cdot$ links to.

To measure the information novelty of one blog, let us first regard each entry in a blog as a document. We first perform Latent Dirichlet Allocation (LDA) [19] to reduce data dimensionality and to generate a topic space for representing entries. We then project each entry onto this topic space to generate a feature vector to represent the entry. After representing each document as a feature vector, cosine similarity or Kullback-Leibler (KL) divergence can be used to calculate the dissimilarity. As demonstrated in [13], cosine similarity is effective on information novelty detection and outperforms KL divergence in several experiments. Thus in this paper, we use the cosine distance to measure information novelty.

Subsequently, the information novelty provided by the hidden node of an entry A_e in blog A is measured as the information novelty between this entry and those entries it links to. The information novelty entry A_e contributes to the network is then calculated as:

$$Nov(A_e|Out(A_e)) = \min_{O_e \in Out(A_e)} Nov(A_e|O_e) \quad (1)$$

Then the information novelty that entry A_e contributes to the network or the hidden node of entry A_e offers is calculated as:

$$Nov(A_e|O_e) = 1 - Cos(A_e, O_e) \quad (2)$$

The information novelty provided by the hidden node of blog A is measured as the average of the novelty scores of the entries it contains.

$$Nov(A|Out(A)) = \frac{\sum_{A_e \in A} \left(Nov(A_e|Out(A_e))\right)}{card\left(Set(A_e)\right)} \quad (3)$$

where A_e is an entry of interest in blog A, and $card\left(Set(A_e)\right)$ is the total number of entries of interest in blog A. In this paper, information novelty ranges strictly between zero and one.

Given the information novelty of each blog, to calculate the InfluenceRank of a blog, let us first regard a set of n blogs as a directed graph with the adjacency matrix $\mathbf{G}$. $G_{ij} = 1$ if blog i links to blog j and $G_{ij} = 0$ otherwise. Then we scale the adjacency matrix $\mathbf{G}$ by its row sums to obtain a normalized adjacency matrix $\mathbf{W}$. The InfluenceRank of a blog A, denoted as $IR(A)$, indicates the opinion leadership of A. InfluenceRank is calculated as the combined opinion leadership of its neighbors including the hidden nodes:

$$IR(A) = (1-\beta) \sum_{i \in In(A)} IR(i) \cdot W_{iA} + \beta \cdot Nov(A|Out(A)) \quad (4)$$

where parameter β is used to adjust how important the information novelty is for the leadership, and $In(\cdot)$ denotes the set of nodes that node $\cdot$ is linked to.

Generally speaking,

$$\mathbf{IR}^T = (1-\beta) \cdot \mathbf{IR}^T \cdot \mathbf{W} + \beta \cdot \mathbf{Nov}^T \quad (5)$$

InfluenceRank can be fitted into a random walk framework which consists of nodes, hidden nodes, and the links among them. The parameter β reflects how significant the novelty is to the opinion leaders we expect to detect. When β is large, we tend to find those opinion leaders who contribute more novel information. An

extreme case is when $\beta = 1$, $\mathbf{IR}^T = \mathrm{Nov}^T$, in which we favor the most novel information contributors to be detected as opinion leaders. When β is small, the importance of the blog in the network is more significant to opinion leaders we expect to detect. Take an extreme $\beta = 0$, InfluenceRank reduces to PageRank.

Due to the presence of dangling nodes (nodes that do not have any out-links) and cyclic paths in the network, the unique solution of Eq. (5) may not exist. To remedy this problem, we apply the remedy of "random jumps" as what PageRank does [9]. The normalized adjacency matrix $\bar{\mathbf{W}}$ after the adjustment can be written as:

$$\bar{\mathbf{W}} = \alpha \mathbf{W} + \left(\alpha \mathbf{a} + (1-\alpha)\mathbf{e}\right)\mathbf{e}^T \tag{6}$$

where α is the probability that the random walk follows a link, $1-\alpha$ is the probability of a "random jump", and $\delta = (1-\alpha)/n$ is the probability that a particular random page is chosen to make this random jump. A typical value is $\alpha = 0.85$. $\mathbf{e}$ is the n-vector of all ones and $\mathbf{a}$ is the vector with components $a_i = 1$ if ith row of $\mathbf{W}$ corresponds to a dangling node, and 0, otherwise. Thus, we have:

$$\begin{aligned} &\mathbf{IR}^T(\mathbf{I} - (1-\beta)\bar{\mathbf{W}}) = \beta \cdot \mathbf{Nov}^T \\ &\mathbf{IR}^T(\mathbf{I} - (1-\beta)(\alpha \mathbf{W} + \left(\alpha \mathbf{a} + (1-\alpha)\mathbf{e}\right)\mathbf{e}^T)) = \beta \cdot \mathbf{Nov}^T \end{aligned} \tag{7}$$

By taking $\mathbf{IR}^T \cdot \mathbf{e} = 1$ into account, Eq. (7) can be written as

$$\mathbf{IR}^T(\mathbf{I} - (1-\beta)\alpha \mathbf{W} - (1-\beta)\alpha \mathbf{a} \cdot \mathbf{e}^T) = (1-\beta)(1-\alpha)\mathbf{e}^T + \beta \cdot \mathbf{Nov}^T \tag{8}$$

with $\mathbf{IR}^T \cdot \mathbf{e} = 1$

Thus, InfluenceRank can be obtained by solving the linear system in Eq. (8). A variety of numeric methods including Jacobi, Gauss-Seidel, or Conjugate Gradient could be used to solve this linear system [18].

Figure 2 summarizes the proposed InfluenceRank algorithm, where we identify opinion leaders by taking into account how important the node is in the network and how novel the information it contributes to the network.

Algorithm: InfluenceRank algorithm
Input: **G**: adjacency matrix,
content in each node (either a blog or an entry)
Output: Detected opinion leaders with InfluenceRank scores
Begin
1) Generate **W** by normalizing **G**
2) Generate topic vectors to represent the content of nodes by LDA
3) Calculate the information novelty for each node by either Eq. (3) (for blogs), or Eq. (1) (for entries)
4) Compute InfluenceRank by solving Eq. (8)
End

Figure 2: The InfluenceRank Algorithm

4. EXPERIMENTS

In this section, we evaluate the performance of our InfluenceRank algorithm on a blog dataset. We first describe the dataset and the metrics we shall use for evaluation. Then we demonstrate the effectiveness of our proposed algorithm on identifying opinion leaders.

4.1 Experimental Setup

4.1.1 Dataset

The dataset we use for experimental studies is collected from an NEC focused blog crawler. The NEC crawler starts by crawling a manually-selected set of seed blogs with technology focus. Then the blogs that are densely connected to the seed blogs are added to expand the seed blogs. The crawler continuously monitors the seed blogs for new entries over time. The data we use in this paper are crawled between July 2005 and October 2006. We clean the crawled data by first removing stop words from entries and then removing entries that contain less than ten terms. The cleaned dataset contains 407 English blogs with 67,549 entries and with 11,187 key terms. We further extract links from the entries. After removing the self links (a link is a self link if it points to the same blog that it belongs to), we extract 12,383 entry-to-entry links.

4.1.2 Evaluation Metrics

To evaluate the performance of our InfluenceRank algorithm from perspectives of link structure and content, we introduce metrics of *coverage*, *diversity*, and *distortion*. Since opinion leaders directly affect the tipping of an innovation, given a set of opinion leaders detected by an algorithm, the number of nodes they could "influence" in the adoption network should be a good indicator of how good the detected opinion leaders are. We introduce the concept of *coverage* due to this intuition.

Definition (Coverage): Given a set of nodes in a network, the *coverage* is defined as the number of nodes that are either directly or indirectly influenced by this set of nodes. Based on the concept of coverage, we further define two metrics that measure the direct and indirect influence of a set of nodes.

Metric 1. One-Step Coverage: Given a set of nodes in a network, one-step coverage is defined as the number of nodes that are *directly* influenced by this set of nodes. In the blogosphere, the one-step coverage is measured as how many blogs directly link to this set of blogs. When the number of nodes in the set equals to one, the one-step coverage is reduced to the number of in-links this node has.

Metric 2. All-Path Coverage: Given a set of nodes in a network, all-path coverage is defined as the number of nodes that are either *directly* or *indirectly* influenced by this set of nodes. In the blogosphere, the all-path coverage is measured as how many blogs can reach this set of blogs by following any path of links.

From the content perspective, as what we demonstrated in the motivating example in Figure 1, the top opinion leaders should be diversified and complementary. Opinion leaders should be able to represent different perspectives of the original information space. The identified opinion leaders are also samples from the original information space in a way. The topic distribution over the opinion leaders should be consistent with that of the original information space. Due to the intuition, we introduce the metrics of *diversity* and *distortion*.

Metric 3. Diversity: The diversity of a set of items is defined as the average pairwise dissimilarity of the items $v_i, i = 1, 2, ..., n$.

$$Diversity(v_1, v_2, ..., v_n) = \sum_i \sum_j \left(1 - Cos(v_i, v_j)\right) \Big/ \left(n \cdot (n-1)/2\right) \tag{9}$$

Metric 4. Distortion: Given the topic distributions of the original information space p_o and the sampling space p_s, the distortion of the sampling space over the original space is defined as the KL divergence of p_s and p_o:

$$D_{KL}(p_o, p_s) = \sum_x p_o(x) \cdot \log\left(p_o(x)/p_s(x)\right) \tag{10}$$

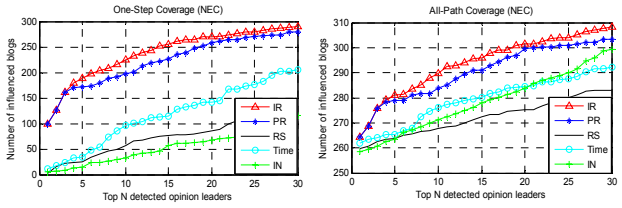
In this paper, the sampling space is composed of top opinion leading blogs.

To illustrate the performance of our proposed InfluenceRank in terms of identifying opinion leaders from perspectives of content and link structure, we compare the performance of our InfluenceRank algorithm (short as *IR*) with *1) PageRank (PR); 2) Random Sampling (RS); 3) Time-based Ranking (Time); 4)Information Novelty-based Ranking (IN).* Specifically, PageRank is purely based on the link structure; Random Sampling selects the

opinion leaders entirely by chance; Time-based Ranking orders blogs by the time they publish posts; and Information Novelty-based Ranking selects opinion leaders based on the information novelty scores of the blogs, which are calculated by the method described in Section 3.

4.2 Identifying Opinion Leaders

In this subsection, we compare the performance of the opinion leaders identified by five algorithms. We first demonstrate the coverage of the opinion leaders identified by five algorithms. Figure 3 illustrates how the one-step and all-path coverage change with the number of identified opinion leaders. The results show that InfluenceRank consistently achieves better one-step and all-path coverage comparing to four baseline algorithms.



(a) One-Step Coverage (b) All-Path Coverage

Figure 3: Coverage comparison for opinion leaders

We then compare the diversity of opinion leaders identified by five algorithms in Figure 4. The opinion leading blogs identified by InfluenceRank tend to provide higher diversity at the higher ranked opinion leaders comparing to four baseline algorithms. Finally we compare the distortion of the opinion leaders identified by five algorithms over the original information space in Figure 5. We use the distribution of two topics ($K = 2$): political and technological (*93:314*) as the topic distribution over the original information space. Among five algorithms, InfluenceRank consistently achieves relatively low distortion.

In summary, in NEC blog dataset, the opinion leaders identified by InfluenceRank achieve better performance in terms of coverage, diversity, and distortion comparing to four baseline algorithms. Due to lack of space, we have not been able to give the results regarding different queries.

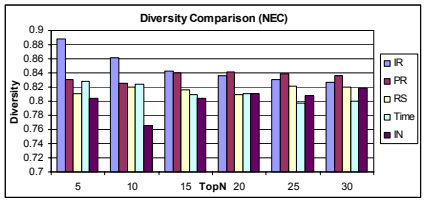


Figure 4: Diversity comparison for opinion leaders

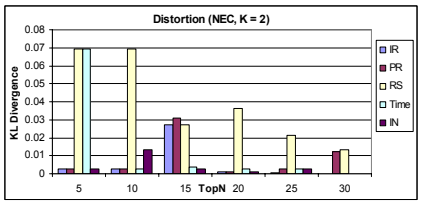


Figure 5: Distortion comparison for opinion leaders

5. CONCLUSIONS

The blogosphere is a fruitful media to understand people's opinions. However, due to its great amount of content and complex link structure, the blogosphere is difficult to be fully understood. Opinion leaders are those who bring in new information, ideas, and opinions, then spread them down to the masses. They are the most informative and influential nodes, and capture the most representative opinions in the network. Identifying opinion leaders is therefore important for understanding the massive and complex blogosphere.

In this paper, we propose a novel ranking algorithm, InfluenceRank, to identify those opinion leaders who are novel information contributors and also highly influential in the network. In the proposed algorithm, we define a way to measure the information novelty of entries or blogs. Then InfluenceRank is proposed to identify opinion leaders by taking into account the importance and the information novelty of nodes in the network. Furthermore, the solution to the InfluenceRank is also provided. Experiments on NEC blog dataset demonstrate that the opinion leaders detected by InfluenceRank achieve better performance in terms of coverage, diversity, and distortion comparing to four baseline algorithms.

6. REFERENCES

[1] R. B. Cialdini, Influence: Science and Practice, Apr 2003.

[2] E. Katz and P. Lazarsfeld, Personal Influence, New York: The Free Press, 1955

[3] E. M. Rogers, Diffusion of Innovations, The Free Press: New York, 1995.

[4] P. Domingos and M. Richardson, Mining the Network Value of Customers, KDD 2001.

[5] D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network, KDD, 2003

[6] X. Song, B. L. Tseng, C.-Y. Lin, M.-T. Sun: Personalized recommendation driven by information flow. SIGIR, 2006

[7] X. Song, Y. Chi, K. Hino, and B. L. Tseng, Information Flow Modeling based on Diffusion Rate for Prediction and Ranking. WWW, 2007

[8] M. R. Solomon, Consumer Behavior. Needham Heights, MA. Allyn & Bacon. 1992.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107-117, 1998.

[10] K. Fujimura, T. Inoue, and M. Sugisaki. The eigenrumor algorithm for ranking blogs. Annual Workshop on the Weblogging Ecosystem, 2005.

[11] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604-632, 1999.

[12] J. Allan, V. Lavrenko, and H. Jin, First story detection in TDT is hard, Proceedings of CIKM, pp. 374-181, 2000.

[13] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering, SIGIR, 2002.

[14] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. "Information Diffusion Through Blogspace," WWW 2004, New York, May 2004.

[15] E. Adar, and L. A. Adamic, Tracking Information Epidemics in Blogspace, WI, pp. 207-214, 2005.

[16] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the Bursty Evolution of Blogspace, WWW, Budapest, Hungary, May 2003.

[17] S. Nakajima, J. Tatemura, Y. Hino, Y. Hara, and K. Tanaka. Discovering important bloggers based on analyzing blog threads. Annual Workshop on the Weblogging Ecosystem, 2005.

[18] G. M. D. Corso, A. Gullí, and F. Romani, Fast PageRank Computation via a Sparse Linear System, *Internet Math.* 2(3), 251–273, 2005.

[19] D. Blei, A. Ng, and M. Jordan, Latent Dirichlet allocation, *J.of Machine Learning Research*, 3:993-1022, Jan 2003.