MASS: a Multi-fAcet domain-Specific influential blogger mining System

Yichuan Cai, Yi Chen

Arizona State University
{yichuan.cai,yi}@asu.edu

***Abstract*— With rapid development of web 2.0 technology and e-business, bloggers play significant roles in the blogosphere as well as the external world. In particular, influential bloggers can bring great business values to modern enterprise. Despite that several systems for mining influential bloggers are available, they measure the influence of bloggers in general rather than domain specific, which is not applicable for real application requirements, such as business advertisement, personalized recommendation and so on. In this paper, we propose an effective model to mine the top-k influential bloggers according to their interest domains, the impact and attitude of the comments to their posts, as well as their authority in the network of page links. In this demonstration, we present MASS, an effective system for mining influential bloggers. We will present the techniques in MASS, its experimental evaluation, as well as its applications.**

I. INTRODUCTION

Web 2.0 provides the second generation web-based communities with services such as forums, wikis, blogs, folksonomies, etc, which can facilitate the communication, collaboration, and information sharing among web users. Blogs, as one of the most important components of web 2.0 service, provide a conductive platform for web bloggers to post their logs of events and to share their personal insights with others.

In light of that blog readers are likely to be influenced by the bloggers, an increasing number of corporations now start to use blogs as a new product marketing strategy to enlarge their profits. Identifying influential bloggers can bring potentially large business opportunities for two reasons. First, the posts from an influential blogger can have a large impact on the purchasing decision of their readers compared with companies' advertisements, as people typically trust and act on recommendations from knowledgeable people and their friends. Second, communication and analysis of influential bloggers bring more insight of the key concerns and new trends of customers' interest on products with much less cost compared with searching, aggregating and analyzing all the blogs.

Recently the topic of identifying top-k influential bloggers begins to attract more and more interests in research community [1], [2]. They measure the influences among bloggers based on "post-reply" relationships, modeled in an *influence graph*. Figure 1 is a sample *influence graph*, in which `Amery` has two posts, $post_1$ with comments from `Bob` and `Cary`, which discusses some programming skills in computer science, and $post_2$ with comments from `Cary` which investigates the recent economic depression and possible trends in the next couple of months.

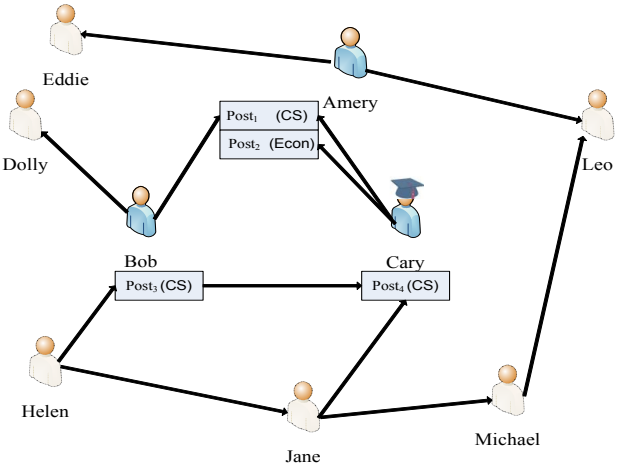

Fig. 1. A Sample of Influence Graph

To evaluate the influence of `Amery`, existing system [1] considers factors such as inlinks/outlinks of the posts, the number of comments as well as the length of the comments.

Although it is intuitive, some valuable information embedded in the "post-reply" relationship is remained to be captured. First, `Amery` have posts in different domains, and her influence power can differ among the domains. That is, the influence of a blogger should be measured in a *domain specific* manner. Second, different commenters can indicate different influence power of a post. For instance, comments from an expert may indicate larger influence of the post compared with comments from a lay user. Thus *citation factor* must be considered. Third, the comments can be positive, negative or neutral. The *attitude* of the commenter should be considered when evaluating the post's influence. Furthermore, besides the "post-reply" relationship, the influence of a post is also reflected in the links to it. For example, when a person finds a blog interesting, s/he may directly add a link to it in her/his own space. External links to a blog provides another metrics to measure the influence of the blogger, like PageRank [3] and HITS [4]. We also take this *authority factor* into consideration when we evaluate the influence of a blogger.

Our system, MASS, can capture the domain specific, citation, attitude and authority information in order to better understand the "post-reply" relationship when we mine the most influential bloggers. The experiment evaluation shows

MASS's effectiveness. More technical details can be found at [5].

II. MULTI-FACET DOMAIN SPECIFIC INFLUENTIAL BLOGGERS

We now briefly describe the data model and define the problem of finding influential bloggers. Given a set of bloggers with their posts, the comments on the posts and the corresponding commenters, and a set of predefined domain categories, we find out the top-k most influential bloggers on each domain.

A blogger's influence on a specific domain is a component of his/her overall influence, hence we first quantify each blogger's overall influence. Intuitively, an influential blogger has high quality posts with many comments, and high authority in the network. The posts and comments reflect a blogger's expertise and popularity, while the authority reflects his/her position in the network. Thus the overall influence of a blogger should consist of two parts: the summation of his/her posts' influence, denoted as `Accumulated Post (AP)` influence score, and his/her authority in the network, noted as `General Links (GL)` influence score. Since each post is domain specific, we choose "post" as the analysis unit, rather than a blogger. The GL is similar to a webpage authority and PageRank. We define the overall influence of a blogger as following:

$$Inf(b_i) = \alpha * AP(b_i) + (1-\alpha) * GL(b_i) AP(b_i) = \sum_{k=1}^{|P(b_i)|} Inf(b_i, d_k) \tag{1}$$

where $GL(b_i)$ is the GL score of blogger b_i, α is the parameter to tune the relative importance of AP score and GL score, which is set to 0.5 as the default value in our system. Furthermore, $|P(b_i)|$ is the total number of posts written by b_i, and $Inf(b_i, d_k)$ is influence score of blogger b_i's post d_k (which will be discussed later).

To define the score of a post d_k of blogger b_i, $Inf(b_i, d_k)$ in Eq. 1, we consider both the quality of the post's content and the commenters' impact.

$$\begin{aligned} &Inf(b_i, d_k) = \beta * QualityScore(b_i, d_k) + \\ &(1-\beta) * CommentScore(b_i, d_k) \end{aligned} \tag{2}$$

where β is a parameter used as a weight for the two parts, set to 0.6 according to empirical study.

$QualityScore(b_i, d_k)$, the first component of $Inf(b_i, d_k)$, is evaluated by the length of a post according to existing work [1]. The longer a post, the higher quality it is considered. When measuring the quality of a post, we also consider its novelty $Novelty(b_i, d_k)$, i.e. whether it is an original idea or a carbon copy from others. As pointed out by [2], reproduced content usually brings little influence to readers. We collect a set of words indicating that an article is a copy of other sources, and set $Novelty(b_i, d_k)$ to a value between 0 and 0.1 if the article contains such words, and otherwise we consider the article original and set its $Novelty(b_i, d_k)$ to 1. We measure $QualityScore(b_i, d_k)$ as the product of a post's length and its novelty.

$CommentScore(b_i, d_k)$, the second component of a post's influence, measures the score of a post by the comments it receives. Each comment's $CommentScore$ is proportional to the summation of the commenter (b_j)'s overall influence score $Inf(b_j)$ and b_j's attitude toward b_i's post d_k, which is the sentiment factor $SF(b_i, d_k, b_j)$. Also, one commenter may put multiple comments on other blogger's posts, and his/her impact to peers should be shared. Hence we normalize the comment score by the total number of comments $TC(b_j)$ of commenter b_j. Note that our work differentiate the importance of the commenters rather than just counting the number of comments as used in [1]. The $CommentScore$ is defined as following:

$$CommentScore(b_i, d_k) = \sum_{j=1}^{|C(b_i,d_k)|} \frac{Inf(b_j) * SF(b_i, d_k, b_j)}{TC(b_j)} \tag{3}$$

$|C(b_i, d_k)|$ represents the total number of commenters who have commented on blogger b_i's post d_k.

The sentiment factor $SF(b_i, d_k, b_j)$, which captures the commenter's attitude, can be classified into three categories: positive, negative or neutral. We assign $SF(b_i, d_k, b_j)$=1 for positive comments (which contain positive words such as "agree", "support", "conform") and set $SF(b_i, d_k, b_j)$=0.1 for negative comments, and $SF(b_i, d_k, b_j)$= 0.5 otherwise.

From Eq.2 and Eq. 3, we get the following equation:

$$\begin{aligned} &Inf(b_i, d_k) = \beta * QualityScore(b_i, d_k) + (1-\beta) * \\ &\sum_{j=1}^{|C(b_i,d_k)|} \frac{Inf(b_j) * SF(b_i, d_k, b_j)}{TC(b_j)} \end{aligned} \tag{4}$$

A blogger's total influence score $Inf(b_i)$ is the summation of the influence score of each post $Inf(b_i, d_k)$ that b_i makes.

Now we evaluate a blogger's influence score for each domain. Intuitively, a post d_k (by blogger b_i)'s influential score with respect to a specific domain C_t is proportional to the post's total influence score and the possibility $iv(b_i, d_k, C_t)$ of d_k belonging to C_t. Then b_i's influence score in domain C_t is the summation of the influence scores of all b_i's posts, where $|P(b_i)|$ denotes the total number of posts by b_i.

$$Inf(b_i, C_t) = \sum_{k=1}^{|P(b_i)|} Inf(b_i, d_k) * iv(b_i, d_k, C_t) \tag{5}$$

The domains can be predefined by the business applications or automatically discovered using existing topic discovery techniques [6]. Given a set of domains, MASS automatically analyzes the posts and generates a $iv(b_i, d_k, C_t)$ using naive Bayesian method [7]. Other interests mining methods [8], [9] can also be plugged into our system.

The vector of b_i's influence scores on every domain is denoted as $Inf(b_i, IV)$.

Application Scenarios

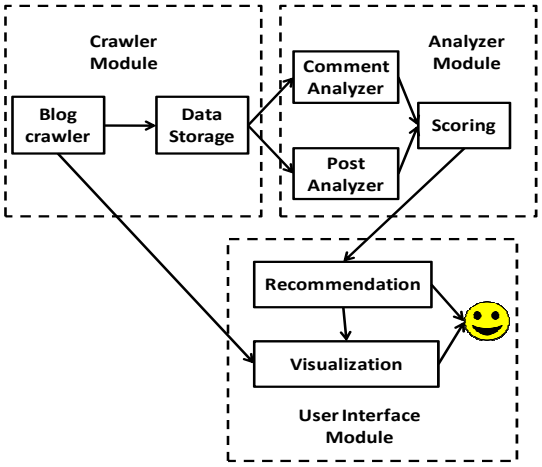

Fig. 2. System Architecture of MASS

With the bloggers' domain influence scores, we can make a recommendation to both business applications and personalized recommendation. There are two application scenarios in our system:

`Scenario 1: Business Advertisement.` As we discussed in Section I, many business corporations realize the value of influential bloggers and start to send advertisements to them. Our system can find out the top-k most influential bloggers in the domains that correspond to the interest domains in business advertisement.

We first mine the interest vector from a user-input advertisement a_l, denoted as $iv(a_l)$. Then we compute the dot product of the vector of a blogger b_i's domain influence scores $Inf(b_i, IV)$ and the interest vector of the advertisement $iv(a_l)$ as blogger b_i's influence with respect to the domains indicated by a_l, denoted as $Inf(b_i, a_l)$. Our system will recommend top-k influential bloggers based on $Inf(b_i, a_l)$ to the corporation.

`Scenario 2: Personalized Recommendation.` Recommendation in social network is popular. For instance, Facebook recommends friends' friends as "the people you may know". MASS recommends influential bloggers considering the domains that the current user is interested in.

III. SYSTEM ARCHITECTURE AND EVALUATION

System Architecture:

The system architecture of MASS is presented in the Figure 2, which consists of three modules: the crawler, the analyzer, and the user interface.

The *Crawler Module* uses a multi-thread crawling technique to efficiently crawl blogosphere and stores the bloggers' information (including the bloggers' personal information, posts, and corresponding comments) in XML files.

The *Analyzer Module* is composed of two subparts: *Post Analyzer* uses text classification technique to classify a post into different domains, and *Comment Analyzer* uses the techniques discussed in Section II to calculate the domain influence score for each blogger.

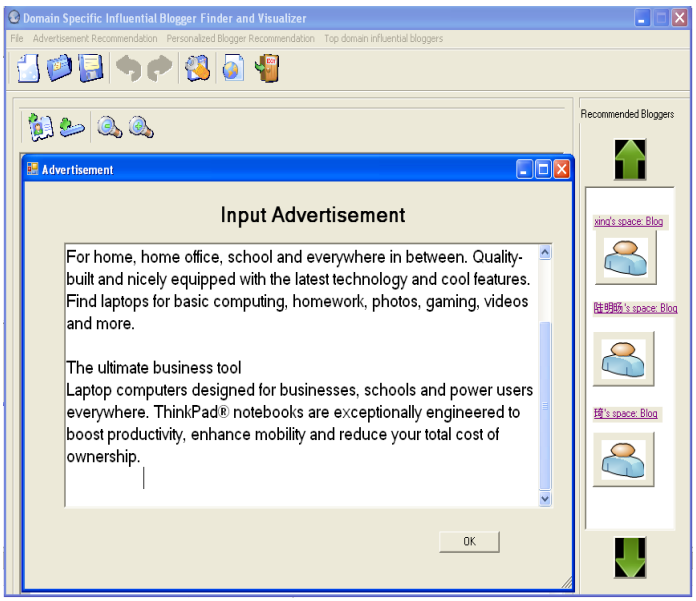

Fig. 3. Advertisement Input Function for MASS

The *User Interface Module* not only provides the recommendation service, but also provides user-friendly panel to visualize the bloggers' network.

Evaluation: We use Microsoft MSN space as test data set, which is one of the most popular blog service providers. Each blogger can write posts on their own blogs and leave comments on others' posts. We have crawled around 3000 MSN spaces with user profiles, comments and about 40000 recent posts. We predefine ten interest domains: {Travel, Computer, Communication, Education, Economics, Military, Sports, Medicine, Art, Politics}.

To evaluate the effectiveness of MASS, we invite 10 users who are graduate student and always write blogs to do a user study, who compare the recommendation performance of top 3 influential bloggers mined from general domain and specific domains. For the top 3 bloggers in the general and domain-specific list, we send the URL of each blogger to the end users, and ask users to score them from 1 to 5 according to their understanding of a specific application scenario, e.g. "Suppose you are the sales manager in Nike, which blogger will you choose to send advertisement to? ". We also compared with Microsoft Live Index [10], which is based on traditional link analysis. The average scores of these systems obtained from the user study, over Travel, Art and Sports domains, are shown in Table I. As we can see, MASSis preferred by users over general influential blogger recommendation and Microsoft Live Index.

IV. DEMONSTRATION

What will be shown in demo To use MASS, a user can load the blogger data set that is crawled offline. Alternatively, the user can also specify a portion of the blogoshere that s/he is interested in. For instance, the user can specify a seed of

TABLE I

USER EVALUATION OF AVERAGE APPLICABLE SCORES FOR INFLUENTIAL BLOGGERS (GENERAL VS. LIVE INDEX VS. DOMAIN SPECIFIC)

Average Applicable Scores	Travel	Art	Sports
General	3.2	3.2	3.2
Live Index	3.0	3.3	3.1
Domain Specific	4.3	4.1	4.6

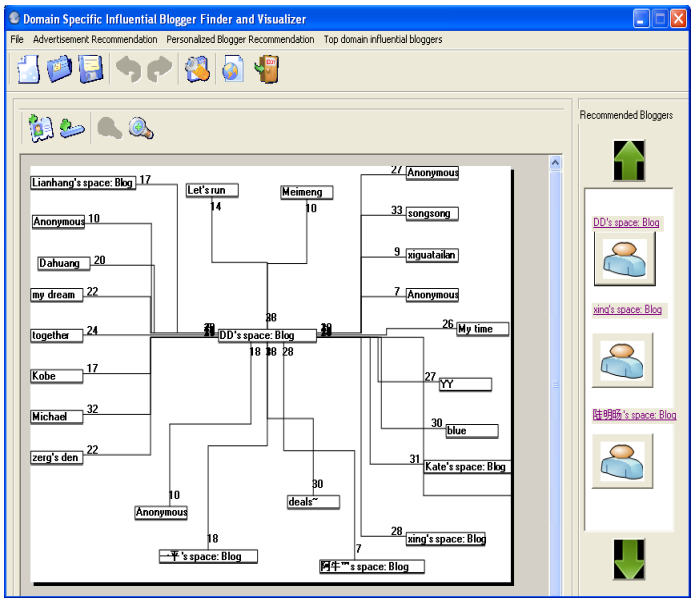

Fig. 4. Post-reply Network of MASS

the crawling (a blogger with a lot of comments and friends, or a valid MSN live space URL), from which the crawling starts. The user can also specify the radius of network where the crawling is performed. In this way, the user can request MASSto find influential bloggers in her/his friend network, rather than the ones in the whole blogosphere.

MASS can support business advertisement application as discussed in Section II. There are two options for a business partner to configure their application requirements: providing advertisement text or choose a domain according to their business applications, which is shown in Figure 3. In the first option, based on the input advertisement, MASS analyzes the content of the advertisement and provides top-k domain-specific bloggers according to the domains mined from the advertisement. In the second option, the business partner selects one or more relevant domains from a dropdown list, e.g. the representative from Nike would choose "Sports" as their category. If no domain is select, MASS can show the top-k bloggers with the largest general domain scores to the user.

MASS can also recommend influential bloggers to bloggers based on their interests, as discussed in Section II. When a new user inputs his/her profile, MASS will extract the domain interest information from the profile and recommend top-k influential bloggers in these domains to the new user. An existing blogger can choose a domain and request MASS to recommend the top-k influential bloggers in this domain.

MASS also allows users to use the toolbar to set personalized parameters for modeling general influence and domain influence as discussed in the section II, such as the relatively importance of a post's quality score and comment score.

From the list of recommended top-k influential bloggers on the right panel of the user interface, the user can double click one to check his/her "post-reply" network in the blogosphere, which will be visualized in the left panel of the user interface. Each node represents one blogger, we will display his/her user name. If a users want to know the detailed influence properties of the blogger (such as the total influence score, domain influence score, the number of posts, the link to important posts, etc.), by double clicking the blogger node she can see all the related information shown in a pop-up window. A line between two nodes represents the post-reply relationship between two bloggers and the number on the line records the total number comments of one blogger on the other blogger's posts, which is shown in Figure 4. The user can easily drag and move nodes in the visualization panel, and zoom in or zoom out the network to get a better view. The visualization graph can be saved as an XML file and be loaded in future.

V. CONCLUSIONS

This paper ventures into a quickly expanding research area: to identify influential bloggers. We present a general framework for influential blogger mining, considering their interest domains, the impact and attitude of the comments to their posts, as well as their authority in the network of page links. We developed the MASS system for identifying top-k influential bloggers, whose effectiveness is demonstrated empirically. MASS can be used in multiple application scenarios, including both business advertisement and personalized recommendation.

VI. ACKNOWLEDGEMENT

This material is based on work partially supported by NSF CAREER award IIS-0845647 and IIS-0915438.

REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *WSDM '08*.
[2] X. Song, Y. Chi, K. Hino, and B. Tseng, "Identifying opinion leaders in the blogosphere," in *CIKM '07*. ACM, 2007, pp. 971–974.
[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Tech. Rep., 1998.
[4] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*.
[5] Y. Cai and Y. Chen, "Mining influential bloggers: from general to domain specific," in *KES ' 09*, 2009(to appear).
[6] X. Li, L. Guo, and Y. E. Zhao, "Tag-based social interest discovery," in *WWW '08*, 2008, pp. 675–684.
[7] "http://en.wikipedia.org/wiki/naivebayesclassifier."
[8] Y. Liu, W. Liu, and C. Jiang, "User interest detection on web pages for building personalized information agent," in *IIT 2006*.
[9] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization."
[10] "Live indexed pages(provided by cubestat)http://www.cubestat.com/."