Mining Influential Bloggers:
from General to Domain Specific

Yichuan Cai, Yi Chen
{yichuan.cai, yi}@asu.edu

Arizona State University P.O. Box 878809, Tempe, AZ 85287 - 8809, USA

Abstract. With rapid development of web 2.0 technology and e-business, bloggers play significant roles in the whole blogosphere as well as the external world. Specially, the most influential bloggers can bring great business values to modern enterprise in multiple ways, by increasing market profits and enlarging business impacts. The bloggers' influences can be deployed only in a specific domain, e.g. computer companies only can utilize the influence bloggers' expertise in computer knowledge, not their expertise in modern art or others. Despite that several influential bloggers mining systems are available, none of them consider the domain specific feature and their evaluations are based on generic influence, 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 and network proximity. We investigate an effective algorithm to evaluate a blogger's influence and develop a domain specific influential blogger mining system. The experiment results show that our system can effectively mine influential bloggers and is applicable to diverse applications.

1 Introduction

With the advantage of the modern technology, Web 2.0 provides a second generation web-based communities with services such as forums, wikis, blogs, folksonomies and 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 services, provide a conductive platform for web bloggers to post their logs of events and share their personal insights with the blog visitors, and let them to read and write down feedbacks. By building such a virtual communities, blogs have attracted great interests from web users, industry, as well as research communities and become one of the most popular and widely used web 2.0 services.

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 by leveraging influential bloggers which have a large population of potential readers. The main reasons come from the following perspectives.

First, the posts from an influential blogger often have a larger impact on their readers' purchasing decision than advertisements from the company, since people typically trust and act on recommendations from knowledgeable people and their friends. To carry out promotional effort, using a "word-of-mouth" advertising - marketing the bloggers with strong influence and leveraging the customers themselves to be unofficial
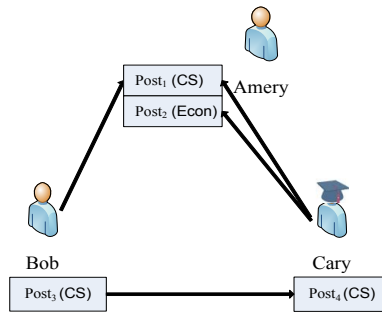
Fig. 1. A Sample "Influence Graph" in Blogsphere

spokesmen, a product can be marketed in a much more cost effective way than traditional methods. Some researchers have given the theoretical explanation of the how the influential members have the influence on the other members' action in the community. [4] investigates causes of user action correlation, which could be categorized into three types: the influence, the homophily and environment. Second, communicating with influential customers and analyzing their blogs can bring companies good understanding and insights of the key concerns and new trends of customers' interest for product improvements with much less cost compared with searching, aggregating and analyzing all the relevant blogs. The influential ones' solutions and suggestions are often very valuable due to the sense of authority they possess.

Due to these potential business opportunities, recently, identifying top-k influential bloggers begins to attract more and more research interests [10, 3, 5]. They measure the influence among bloggers based on the "post-reply" relationships, which are modeled in an *influence graph*.

Let's take a look at a sample *influence graph* in Figure 1, `Amery` has two posts, which are $post_1$ with comments from `Bob` and `Cary`, and $post_2$ with comments from `Cary`. Assume that $post_1$ discusses some programming skills in computer science, while $post_2$ investigates the recent economic depression and possible trends in the next couple of months. To evaluate the influence of `Amery`, existing work [3] considers factors such as the number of inlinks/outlinks, the number of comments as well as the length of those comments, to measure the influence degree of blogger `Amery`.

Although it is a straightforward metric to use, some important and valuable information embedded in the "post-reply" relationship are ignored by the existing works. First, the two posts of `Amery` are related to different domains, when evaluating the influence of `Amery` on computer science or economics, it is necessary to consider them separately. In another word, the influence of a blogger is domain specific, consequently a good model should capture this information. Secondly, the influence of each comment may have different impact power, depending on who issues it. For example, `Cary` is an expert in computer science, while `Bob` is an entry level freshman of computer science. Their comments on `Amery`'s $post_1$ should be treated differently, and it is easy to see `Cary`'s comment would enhance the influence of $post_1$ more. Thirdly, the comments from other bloggers could be positive, negative or neutral, and these sentimental factors also affect the post's influence among commenters.

Furthermore, to evaluate the influence of a blogger, only considering "post-reply" relationship is not enough since people may not put comments on others' blogs even s/he has great interest in it. Also, an influence blogger would usually have many external links to his/her blog, e.g.when people find someone's blog is very interesting, s/he may

directly add a link to his/her own space. External links to a blogger provides another possible way to measure the influence of that blogger. PageRank [8] and HITS [6] are effective algorithms to evaluate the authority of a page (and link). We also take the authority factors into consideration when we evaluate the influence of the bloggers.

In this paper, we propose a mechanism that can investigate the top-k domain specific influential bloggers. Considering both "post-reply" relationship and general link information of a blogger, it is possible to evaluate the comprehensive influence of a blogger.

There are several challenges in the blogger's domain specific influence evaluation. For instance, what should be considered when evaluating a blogger's influence in a specific domain? Should his topology proximity be included in the evaluation? How to capture the quality of a post, and measure its quality? How to judge the degree of the impact of a blogger on a commenter? How to evaluate a blogger's influence on different domains? In the following sections, we investigate these issues in turn and explore our model and system in detail.

The rest of paper is organized as follows: we introduce our interest vector model and discuss how to evaluate the domain specific influence in Section 2. Experimental studies are presented in Section 3 and finally Section 4 concludes the paper.

2 Our Approach

In this section, we will first introduce an interest vector model which represents the domain specific interest for a blogger, and then we discuss how to evaluate the blogger's domain specific influence in detail. At last, we will illustrate how to apply our model in different applications accordingly.

2.1 Interest Vector Model

A blogger's blog space is composed of multiple posts, and each of them can belong to one or multiple possible domains, e.g. some posts of latest NBA news are belong to the `sports` domain, the broadcast of U.S. present Obama's speech on ASU's commencement belongs to the `politics` domain as well as the `education` domain. In order to describe a blogger's interest on a specific domain, we define an "interest vector model", which represents the possibility of a post belonging to a specific domain, as follows:

Given a blogger b_i's post d_k, its interest can be quantified as a vector in the whole interest space, called *interest vector*:

$$IV(b_i, d_k) = \{iv_1, iv_2, ..., iv_N\}$$

where $iv_t \in [0, 1], (1 \le t \le N)$. t is the tth dimension in the interest space, and N is the total number of domains, which can be predefined according to some standard categories(such as Open Directory Project) [1].

Most blog service providers can support predefined categories to users, e.g. MSN Space allows bloggers to select a category from a candidate list for his/her new post, we denote them as "category tag" for the post. However, the coverage of "category tag" is

---
[1] http://www.dmoz.org/

pretty low: according to our sample of 1000 bloggers in MSN spaces in the empirical study, we find about 70% of the posts do not explicitly provide their "category tag".

When a tag information is not available, we mine the blogger's interest information from their post content and construct the *interest vector*. Then we use the naïve Bayesian classifier [2] to get the possibility that a post belongs to certain predefined category. Formally, a post's interest vector is calculated as follows:

$$iv(b_i, d_k, C_t) = \frac{P(b_i, d_k | C_t) P(C_t)}{\sum_{n=1}^{N} (P(b_i, d_k | C_n) P(C_n))} \tag{1}$$

where $P(C_t | b_i, d_k)$ is the possibility that blogger b_i's post d_k belong to category C_t. $iv(b_i, d_k, C_t)$ is the possibility that the blogger b_i's post d_k belong to category C_t normalized by the summation of the possibility that blogger b_i's post d_k belongs to all the predefined categories.

2.2 Influence Evaluation

A blogger's influence on a specific domain can be treated as one component of his/her overall influence, hence before we investigate the domain specific influence for each blogger, it is necessary to quantify each blogger's personal overall influence. Intuitively, an influential blogger always has high quality posts together with many commenters, and high authority in the whole network. The posts and comments reflect a blogger's expertise and popularity, while the authority reflects his/her position in the whole network and linkage with other bloggers. Correspondingly, the overall influence of a blogger can be measured by two parts: the summation of his/her posts' influence noted as `Accumulated Post (AP)` influence score, and his/her authority in the network, noted as `General Links (GL)` influence score. Since each of his/her post reveals his/her interest in specific domains, we choose "post" as our basic analysis unit, rather than a blogger. The GL is similar to a webpage authority and PageRank [8] value becomes a natural choice of the approximation of the authority. The GL score of each blogger can be calculated by standard Page Rank algorithm or direct approximated by certain Page Rank Value provider.[2] Hence, we can define the personal overall influence of a blogger as following:

$$Inf(b_i) = \alpha * \sum_{k=1}^{|P_i|} Inf(b_i, d_k) + (1 - \alpha) * GL(b_i) \tag{2}$$

where $\sum_{k=1}^{|P_i|} Inf(b_i, d_k)$ is AP score, and $|P_i|$ is b_i's total number of posts. Specifically, $Inf(b_i, d_k)$ is influence score of blogger b_i's post d_k (which will be introduced later). $GL(b_i)$ is the GL score, α is the parameter to tune the relative importance of AP score and GL score.

The post's influence score is always reflected by its quality and the comments on the post. In order to define the score of each individual post d_k of blogger b_i, $Inf(b_i, d_k)$ in Eq. 2, we consider both the quality of the post's content and the commenters' impact.

$$Inf(b_i, d_k) = \beta * QualityScore(b_i, d_k) + (1 - \beta) * CommentScore(b_i, d_k) \tag{3}$$

---

[2] "Cubestat (http://www.cubestat.com)"

where β is a parameter used as a weight for the two parts.

$QualityScore(b_i, d_k)$, as the first component of $Inf(b_i, d_k)$, is evaluated by the length of a post in existing works [3]. The longer, the more influence the post has. Besides, the novelty plays an important role. The more novel the post is, the more influence it has. The novelty of a post reflects the creativity of that post, whether b_i's post d_k is his original idea or carbon copy from others. In our setting, the $QualityScore(b_i, d_k)$ is the product of a post's length and novelty. The novelty is a numeric value between 0 and 1, which could be mined from the post content.

$CommentScore(b_i, d_k)$, the second component of a post's influence, reflects the impact on the commenters. Each comment's score, is proportional to the summation of the commenter blogger's personal overall influence score $Inf(b_j)$ and his/her attitude toward the post, which is the sentiment factor $SF(b_i, d_k, b_j)$ of b_j's comment to b_i's post d_k. Also, one commenter may put multiple comments on other blogger's spaces, and his/her impact to peers will be shared, hence we normalize the score by the total comments $TC(b_j)$ of commenter b_j. The $CommentScore$ is defined as following:

$$CommentScore(b_i, d_k) = \sum_{j=1}^{|b_i, d_k|} \frac{Inf(b_j) * SF(b_i, d_k, b_j)}{TC(b_j)} \qquad (4)$$

$|b_i, d_k|$ is the total number of comments on blogger b_i's post d_k. The sentiment factor $SF(b_i, d_k, b_j)$ captures the commenter's attitude, and can be classified into three categories: positive, negative or neutral. We use the following heuristics to predict its value: as long as a comment contains certain positive words (such as "agree", "support", "conform", which are positive comments), we treat it as a positive comment. If it contains negative words (such as "disagree", "hate", "not credible"), we treat it as a positive comment. Otherwise, we treat it as a neutral comment.

From Eq.3 and Eq. 4, we get the following equation:

$$Inf(b_i, d_k) = \beta * QualityScore(b_i, d_k) + (1 - \beta) * \sum_{j=0}^{|b_i, d_k|-1} \frac{Inf(b_j) * SF(b_i, d_k, b_j)}{TC(b_j)} \qquad (5)$$

Each blogger's post has an equation in the format of Eq. 5 , and all the equations of bloggers' post can be solved by iterative method efficiently [1]. The solution to the equation set provides the influence score for each blogger's post, from which we can get each blogger's total influence score by using the Eq. 2.

2.3 Domain Specific Influence Score

Finally, we evaluate the blogger's influence score for each domain. There are several predefined domains which are very popular in the blogosphere, such as Travel, Art, Sports, etc. Intuitively, the post's domain influence score w.r.t certain domain is proportional to the post's total influence score and the $iv(b_i, d_k, C_t)$ (interest score of b_i's post d_k's belongs to predefined category C_t), which is evaluated as following:

$$Inf(b_i, C_t) = \sum_{k=1}^{|P_i|} Inf(b_i, d_k) * iv(b_i, d_k, C_t) \qquad (6)$$

$iv(b_i, d_k, C_t)$ could be calculated by existing interests mining method [7, 9], and we choose naïve Bayesian algorithm [2] in our implementation.

Example 1. Let us consider the influence graph in Figure 1 as our example and explore the whole domain specific influence score calculation in detail. To be concise, we use the first character of blogger's name to represent the blogger, e.g A stands for $Amery$. We can get the following equations for the set of bloggers according to Eq. 5:

$Inf(A, P_1) = \alpha * QualityScore(A, P_1) + \frac{1}{2} * \beta * ((\alpha * Inf(B, P_3) + (1-\alpha) * GL(B)) * SF(A, P_1, B))) + \frac{1}{2} * \beta * ((\alpha * Inf(C, P_4) + (1-\alpha) * GL(B)) * SF(A, P_1, C)))$

$Inf(A, P_2) = \alpha * QualityScore(A, P_2) + \frac{1}{2} * \beta * ((\alpha * Inf(C, P_4) + (1-\alpha) * GL(C)) * SF(A, P_2, C)))$

$Inf(B, P_3) = \alpha * QualityScore(B, P_3)$

$Inf(C, P_4) = \alpha * QualityScore(C, P_4) + \frac{1}{2} * \beta * ((\alpha * Inf(B, P_3) + (1-\alpha) * GL(B)) * SF(C, P_4, B)))$

As we can see from these equations, $Inf(A, P_1)$, $Inf(A, P_2)$, $Inf(B, P_3)$ and $Inf(C, P_4)$ are variables, after we solve these equations, it is easy to get the values of $Inf(A)$, $Inf(B)$ and $Inf(C)$ which are the bloggers overall influence scores. Based on them, we can further approach to domain influence scores $Inf(A, Econ)$, $Inf(B, CS)$ and $Inf(C, CS)$ for domain {Econ, CS} directly.

3 Experiments

In order to evaluate the effectiveness of our method, we conduct comprehensive analysis on real data set in the following subsections. We use Microsoft MSN space [3] as our 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 1000 MSN spaces with user profiles, comments and their recent posts.

We predefine ten interest domains as following: {Traveling, Computer, Communication, Education, Economics, Military, Sports, Medicine, Art, Politics}, because these domains cover most interests of bloggers. When calculate the Genaral Link(GL) influence score, we found that most bloggers' PageRank values are very small. In our data set, 90% blog's PageRank value is less than 1, about 99% blogs' PageRank value is less than 3. Instead of using PageRank directly, we utilize Microsoft Live Indexed Pages as the approximation of PageRank value, which could be obtained from the website "Cubestat"(http://www.cubestat.com/).

3.1 Domain Specific or Not

To evaluate the effectiveness of our model, we invite 10 users to do a user study, who compare the recommendation performance of top 3 influential bloggers mined from general domain and specific domains (Traveling, Art, and Sports). For the top 3 bloggers in the general and domain-specific lists, we send the URL of each blogger to 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 your advertisement to? ". The results of average applicable score for the user study is shown in Table 1. As we can see, our model has better evaluation results than that of general influential blogger recommendation system cross

---

[3] http://home.spaces.live.com/

different domains. Especially, the Sports domain has a much higher evaluation score 4.6 than that of general one of 3.2.

Average Applicable Scores	Traveling	Art	Sports
General	3.2	3.4	3.2
Domain Specific	4.4	4.0	4.6

Table 1. User Evaluation of Average Applicable Scores for Influential Bloggers (General VS. Domain Specific)

3.2 Impact of Weighting Parameters

As we have discussed in Section 2, α is the parameter to tune the related importance of accumulated post influence and general link influence and β is the parameter to adjust the related importance of each post's quality score and comment score. To see the impact of these tuning parameters, we randomly choose the Art domain as an example.We tune parameters α and β by fixing one and changing the other, observe the variance of the ranking results. For space limit reason, we only show sample results in Table 2 with β changing by fixing α=0.5.

$\beta = 0.9$	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.01$	$\beta = 0$
youyou	youyou	youyou	youyou	kelly
sky	sky	sky	kelly	winson
newwishes	newwishes	sabrina	sky	best
sabrina	sabrina	newwishes	sabrina	whenlove
Frank	Frank	kelly	newwishes	youyou

Table 2. Top 5 influential bloggers with different β given α

As we can see from Table 2, the top 1 influential blogger is changed from "youyou" to " kelly". Take a close look at these two bloggers' posts, "youyou" has many high quality posts, together with a lot of positive comments from high influential commenters. While for "kelly", whose posts always reproduce from other sources without rich contents, but have many influential commenters, possibly from her friends on the blogosphere. No matter the quality of her blog's posts are good or not, she always has a high comment score. This example shows the relationship between the two components of a post, and both of them have impacts on the evaluation of a post's influence score.

4 Conclusions and Future Work

In this paper, we address a novel problem of identifying influential bloggers considering domain specific information. To better identify influential bloggers, we analyze the interests of bloggers, evaluate the influence of a blogger according to their interest domains. The evaluation with data from a real world blog site, Microsoft MSN space, shows the effectiveness of our approach.

In the future, we will further extend our system to visualize the influential bloggers and cooperate with real business applications.

5   Acknowledgement

References

1. http://en.wikipedia.org/wiki/gauss-seidelmethod.
2. "http://en.wikipedia.org/wiki/naivebayesclassifier".
3. N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *WSDM*, 2008.
4. A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *sigkdd*, pages 509–516. ACM, 2008.
5. D. Kempe, J. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03*. ACM, 2003.
6. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*.
7. Y. Liu, W. Liu, and C. Jiang. User interest detection on web pages for building personalized information agent. In *Innovations in Information Technology*, 2006.
8. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
9. W. Paik, S. Yilmazel, E. Brown, M. Poulin, S. Dubon, and C. Amice. Applying natural language processing (nlp) based metadata extraction to automatically acquire user preferences. In *Proceedings of the 1st international conference on Knowledge capture*, 2001.
10. J. Scrpps, P.-N. Tan, and Abdol-Hossein. Node roles and community structure in networks. WEBKDD, 2007.