Domain-sensitive Opinion Leader Mining from Online Review Communities

Qingliang Miao, Shu Zhang, Yao Meng, Hao Yu
Fujitsu R&D Center Co., LTD
Dong Si Huan Zhong Rd, Chaoyang District, Beijing P.R. China
{qingliang.miao, zhangshu, mengyao, yu}@cn.fujitsu.com

ABSTRACT

In this paper, we investigate how to identify domain-sensitive opinion leaders in online review communities, and present a model to rank domain-sensitive opinion leaders. To evaluate the effectiveness of the proposed model, we conduct preliminary experiments on a real-world dataset from Amazon.com. Experimental results indicate that the proposed model is effective in identifying domain-sensitive opinion leaders.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Design, Experimentation

Keywords

Opinion leader, Ranking, Social networks

1. INTRODUCTION

The success and popularity of online review communities, such as Amazon.com have stimulated customers to publish online reviews. Customers share their expertise and opinions, provide suggestions, and form groups in online review communities. The customers whose opinions are representative and authoritative are called opinion leaders in the “Diffusion of Innovations Theory”. Usually, opinion leaders reflect the opinions of the masses and have strong influence on other customers. Opinion leader mining has attracted more and more attention and researchers have investigated how to identify opinion leaders from blogosphere, twitters and review communities [1][2][3]. In the field of opinion leader identification, most existing work is focused on identifying opinion leaders at a global level. Since opinion leaders have different expertise and interest, opinion leaders are domain sensitive. Therefore, it is more suitable to identify opinion leaders in domain-level. The key challenge of identifying domain-sensitive opinion leaders is how to model opinion’s authoritativeness and representativeness using customers’ reviews and linking information among customers. In this paper, we propose a ranking model which leverages the link structure of customer networks and their personal expertise and interest.

2. THE APPROACH

Figure 1 illustrates the structure of review communities. Customers publish reviews which belong to different domains. We use domains in which a customer has published reviews as his expertise space. For instance, customer c_1 published two reviews r_1 and r_2 in domain d_1, therefore domain d_1 is one element of c_1’s expertise space.

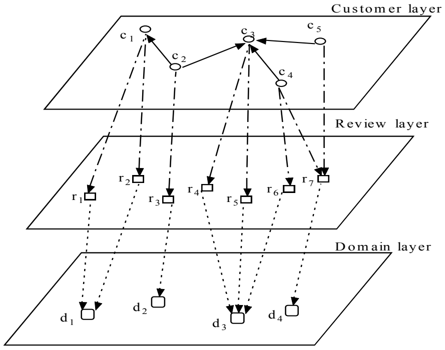

Figure 1. Structure of online review communities.

In practice, customers are interested in others’ opinions and form following relationships. Customers who are interested in others’ opinions are called fans, and customers who are interested by others are called friends. For example, customer c_2, c_4 and c_5 are interested in customer c_3’s opinion, consequently, c_2, c_4 and c_5 are fans of c_3 and c_3 is friend of c_2, c_4 and c_5. We represent the following relationships as directed links. The weights of directed links could reflect different closeness among customers.

Intuitively, the influence of a customer in a customer network is proportional to the combined importance of his fans. In particular, a customer has high influence if the sum of influence of his fans is high. Although the influence of customers and “importance” of web pages shares certain similarities, there are also major differences. Since customers generally have different expertise and interest in various domains, influences of customers also vary in different domains. Consequently, in this paper we identify opinion leaders according to both the link structure and customers’ individual expertise and interest.

Definition 1: Customers’ expertise matrix

Let C be the set of customers, D be the set of domains and V be the matrix which represents customers’ expertise.

Customers set: $C = \{c_1, c_2, c_3, \ldots, c_m\}$;
Domain set: $D = \{d_1, d_2, d_3, \ldots, d_n\}$;
Customers’ expertise matrix: V_{mn};

Each element v_{cidj} in V_{mn} represents the number of reviews published by customer i which belong to domain j. We normalize V according to row as V' and according to column as V''.

We utilize customers’ following relationships to construct a directed graph $G(U,E)$. In G, U represents all the customers and E refers to their relations. Based on the directed graph G, a random surfer model computes as follows: the random surfer visits each customer with certain probability by following the appropriate edge in G. The random surfer performs a domain-sensitive random walk, i.e. the transition probability from one customer to

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

another is domain-sensitive. By doing so, we are essentially constructing a domain-specific graph among customers.

Definition 2: Domain-specific transition probability

Given a domain d, we define the transition matrix P_d, each element in P_d, i.e. the domain-specific transition probability of the random surfer from customer c_i to customer c_j, is defined as follows:

$$t(i,j) = T(i,j) * sim_d(i,j) \tag{1}$$

$$T(i,j) = \begin{cases} 1, & if\ \ i \text{ interests } in\ j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

simd(i, j) represents the expertise similarity between customer c_i and c_j in domain d. *simd(i, j)* can be calculated by the difference of the two customers' expertise in domain d. The more similar the two customers' expertise is, the higher the transition probability from customer c_i to c_j.

$$sim(i,j) = 1 - \left| V'_{id} - V'_{jd} \right| \tag{3}$$

The definition of the domain-specific transition considers two aspects. *T(i,j)* represents the following relationships among customers and *simd(i, j)* captures the customers' expertise and interest. In certain conditions, some customers might be interested in one another and form a close loop. Since customers in this loop are not interested in other customers outside the loop, high influence will be accumulated without distribution. To tackle this issue, we introduce a teleportation vector E_d like [3]. E_d is defined as the *d-th* column of matrix *V''*.

$$E_d = V''_{.d} \tag{4}$$

Based on the domain-specific transition probability matrix P_d and teleportation vector E_d, the customer rank in domain d, CR_d can be computed iteratively by the following formula. Where α is the weight that can be used to adjust the contribution of P_d and E_d.

$$CR_d = \alpha * P_d \times CR_d + (1-\alpha) * E_d \tag{5}$$

3. EXPERIMENTS

In this research, we collect experiment data from Amazon.com. Some key statistics about the dataset are shown in Table 1.

Table 1. Descriptive Statistics of the Dataset

Domain	Product Number	Customer Number
Kitchen	12305	3606
Toy	6061	1786
Sports	3506	1813

To evaluate the approach, we adopt two metrics, coverage and authority as in [3][4]. To evaluate the effectiveness of the approach, we have compared with two baseline algorithms, PageRank and Indegree-based algorithms, as in [2] and [4].

Figure 2 shows how the authority and coverage change with the number of identified opinion leaders by different algorithms. From 2, we can see that our model significantly outperforms the baselines on authority and coverage. This is very important for companies that want to identify opinion leaders with high authority and coverage to accelerate new product diffusion.

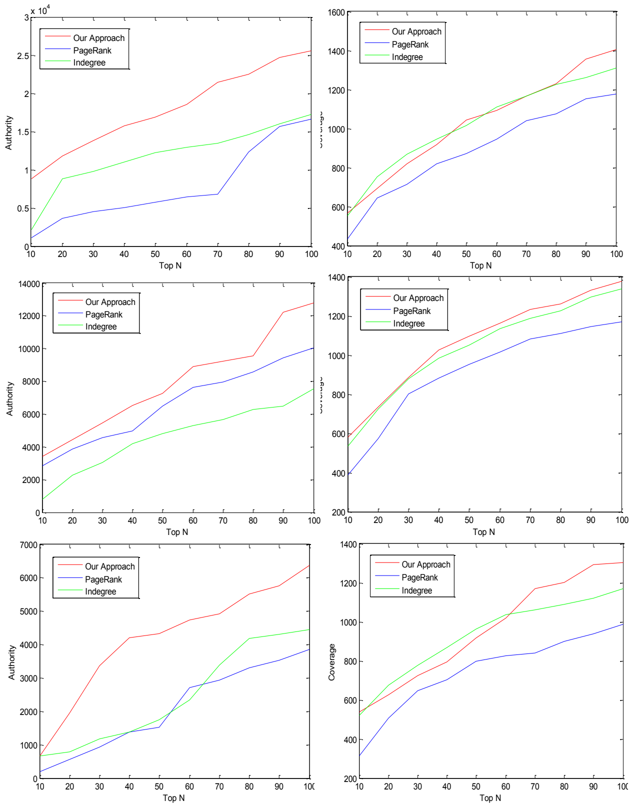

Figure 2. Results on authority and coverage in Kitchen, Toy and Sports domain.

In this study, we investigate how to identify domain-sensitive opinion leaders and proposed a ranking model. Experiments on real world dataset indicate our approach is effective. As a future research, we plan to exploit textual information to discover customers' fine-grained expertise and interest

4. REFERENCES

[1] Song X., Chi Y., Hino K. and Tseng B.L. 2007. Identifying opinion leaders in the blogosphere, In *Proceedings of the Conference on Information and Knowledge Management* (Lisboa, Portugal, November 06 - 09, 2007). ACM, New York, NY, 971-974.

[2] Zhou H., Zeng D. and Zhang C. 2009. Finding leaders from opinion networks, In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*. (Dallas, USA, June 8-11, 2009). IEEE, NJ, USA 266-268.

[3] Weng J., Lim E.P., Jiang J. and He Q. 2010, TwitterRank: Finding topic-sensitive influential twitterers, In *Proceedings of the third ACM International Conference on Web Search and Data Mining*. (New York, USA, February 3-6, 2010). WSDM'10. ACM, New York, NY, 261-270.

[4] Jurczyk P. and Agichtein E., 2007. Discovering authorities in question answer communities by using link analysis, In *Proceedings of the Conference on Information and Knowledge Management*. (Lisboa, Portugal, November 06 - 09, 2007). ACM, New York, NY, 919-922.