Identifying Opinion Leaders from Online Comments

Yi Chen, Xiaolong Wang, Buzhou Tang*, Ruifeng Xu, Bo Yuan, Xin Xiang, and Junzhao Bu

Key Laboratory of Network Oriented Intelligent Computation,
Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China
{chenyi,xuruifeng,yuanbo,xiangxin,bujunzhao}@hitsz.edu.cn,
wangxl@insun.hit.edu.cn, tangbuzhou@gmail.com

Abstract. Online comments are ubiquitous in social media such as micro-blogs, forums and blogs. They provide opinions of reviewers that are useful for understanding social media. Identifying opinion leaders from all reviewers is one of the most important tasks to analysis online comments. Most existing methods to identify opinion leaders only consider positive opinions. Few studies investigate the effect of negative opinions on opinion leader identification. In this paper, we propose a novel method to identify opinion leaders from online comments based on both positive and negative opinions. In this method, we first construct a signed network from online comments, and then design a new model based on PageTrust, called TrustRank, to identify opinion leaders from the signed network. Experimental results on the online comments of a real forum show that the proposed method is competitive with other related state-of-the-art methods.

Keywords: Opinion Leader, Online Comments, Signed Networks, PageRank.

1 Introduction

The social media, such as micro-blogs, forums and blogs, have rapidly developed during the past decades. Online comments provide an important place for reviewers to share their positive or negative opinions toward affairs or products. They have become an important component of social media. Among reviewers, there are several opinion leaders whose opinions greatly affect others. Identifying these opinion leaders from online comments is of great significance. For governments, opinion leaders can help positive opinions toward hot events to be spread rapidly, which will promote social harmony and stability; for enterprises, with the help of opinion leaders, new products can be quickly spread to their customs and achieve a good sale; for publics, knowing opinion leaders means mastering the mainstream viewpoints about hot events or new products.

Many methods have been proposed to identify opinion leaders in social networks. The early methods simply use statistical measurements based on social network analysis, including degree centrality [1], closeness centrality [2], graph centrality [2] and

* Corresponding author.

H.-Y. Huang et al. (Eds.): SMP 2014, CCIS 489, pp. 231–239, 2014.

betweenness centrality [2]. The shortcoming of these statistical measurements is that they may result in finding junk opinion leaders who forge a deluge of links as they only consider network links. Subsequently, a number of relatively complex methods, such as PageRank, HITS [3, 4], TwitterRank [5] and PageRank-like algorithm [6], are proposed for opinion leader identification.

The main limitation of these methods is that they only consider negative opinions, which is not suitable for online comments. Recently, three PageRank-like models are proposed for ranking nodes of networks with negative links, i.e., the Simple PageRank (Sim-PR) [7], Virtual PageRank (Vir-PR) [8] and PageTrust (PT) [9], which are potential to identify opinion leaders from online comments.

In this paper, we propose a novel method to identify opinion leaders from online comments. In this method, we first construct a user network with positive and negative links (called signed network) via four procedures: setting up a basic weight post network with explicit and implicit links, labeling the sign of explicit links, inferring the sign of implicit links, transforming the signed post network into a signed user network. Then we design a new model based on PageTrust, called TrustRank, to identify opinion leaders from the signed network. Compared with other methods, our method considers both positive and negative opinions. In addition, the negative link has two meanings: "negation" sense and "weak-positive" sense. Negation sense means leaving and stopping. Weak-positive sense, for example "the enemy of my enemy is my friends", means keeping and going on. Experimental results on the online comments of a real forum show that the proposed method is competitive with other related state-of-the-art methods.

The remaindering sections of this paper are organized as follows: Section 2 describes how to construct a signed network from online comments. Section 3 introduces a novel model based on PageTrust to identify opinion leaders in the signed network. Section 4 discusses the experiments on an online comment dataset from a real forum. Conclusions are drawn in section 5.

2 Construct Signed Networks

Before illustrating the detail procedures, we define some basic notations. Let $P=\{p_1,p_2,\ldots,p_n\}$ be a comment post set and p_i represents the *ith* post. Let $U=\{u_1,u_2,\ldots,u_m\}$ be a user set and u_j represents the *jth* user. Let u_{p_i} denotes a user who posts p_i . In addition, let $G_P(P,E_P)$ represents a comment post signed network, where E_P denotes the relationship between posts (explicit or implicit). Each edge $p_{ij} \in E_P$ can be expressed by a four-tuple (p_i,p_j,w_{ij},s_{ij}), where w_{ij} denotes the edge weight ranging from (0, 1], s_{ij} denotes the edge sign ranging from {-1, 1}. Similarly, let $G_U(U, E_U)$ represents a user signed network and each edge $u_{ij} \in E_U$ can be denoted by a four-tuple (u_i,u_j,w_{ij},s_{ij}). Note that commonly the number of set P is not tantamount to the set U because some users may have two or more posts.

2.1 Network Construction

Explicit Link. In an online forum, the explicit link is denoted by two meta operations: reply and citation. For posts p_i and p_j, if p_j is a direct reply toward or cite p_i, there exist an explicit link from p_j to p_i . Note that if p_m which cites p_j is a reply to p_i, there is an explicit link from p_m to p_j rather than p_i. And the weight value is 1 for all explicit links.

Implicit Link. In an online forum, p_i and p_j $(1 \le i, j \le n)$ have not an explicit relationship, however, they share some semantic similarities. There exist an implicit link from p_j to p_i because before posting, users commonly have read and also been influenced by several preceding posts. We adopt a method for measuring post similarity [10] to calculate the relevancy between two posts. The implicit link weight is equal to the post similarity.

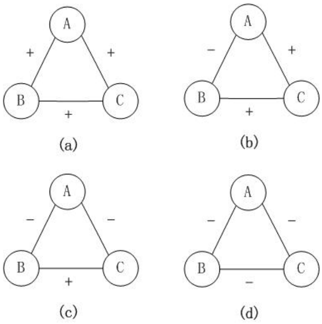


Fig. 1. Illustration of structural balance theory. Triads with odd number of pluses are labeled as balanced (A and B) and Triads with even positive edges are labeled as unbalanced (C and D).

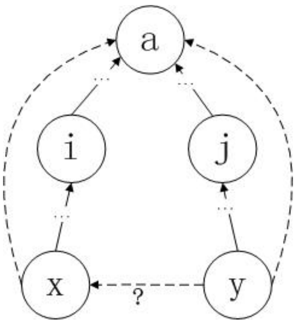


Fig. 2. A virtual triad *axy*. s_{yx} denotes the implicit link sign, s_{xa} and s_{ya} denotes virtual link sign.

2.2 Network Labeling

We classify comment posts into two types: direct and indirect posts. The direct posts show a direct attitude toward the target post author, including “excellent post”, “agreement” or “junk post”; the indirect posts illustrate sentiment opinions toward the role talked in the target post. It is not difficult to differentiate them because the direct posts have several particular characteristics: laconicism; slang; accompanied by special punctuations.

Two strategies are adopted for labeling the direct and indirect posts. For the former, we use two artificial sets collected according to the special characteristics. This is practicable because in a certain media-sharing the direct posts are often enumerable. For the indirect posts, three procedures are necessary for labeling: first setting up a heterogeneous triad for sharing a common object, then utilizing the sentiment analysis to label the direct attitude, at last multiplying the direct sign.

Sentiment Analysis. Firstly, the comment post is split into several sentences. Secondly, six artificial features are extracted and a common classification method is selected for identifying the sentence sentiment orientation. Lastly, the post sentiment orientation is in accordance with the sign of the large number of sentence orientation. All features are set 1 or 0 according to whether they appear or not. The six artificial features include: ①8664 positive words, we collect 2036 positive words manually and use these words as seeds for expending in the Chinese thesaurus set (tongyici cilin) and the final set contains 8664 positive words; ②16894 negative words; ③13 negation words; ④169 degree words; ⑤130 degree words, such as “think”; ⑥16 special punctuations, such as “!!!!!” and “????”.

2.3 Network Inferring

Structural Balance. Structural balance theory is originated in social psychology and then formulated by Heider in the 1940s [11]. Figure 1 shows several triad examples of the structural balance. A simple triad is balanced based on two kinds of situations: the sign of three edges are all positive; there are two negative and one positive signs. They are in accordance with the intuition that "the friend of my friend is my friend" and "the enemy of my enemy is my friend".

For inferring the implicit links labels, it is necessary to build up a virtual triad. As can be seen in figure 2, for inferring the label of implicit link s_{yx}, we need to build up a virtual triad *axy* and also should know the link sign s_{xa} and s_{ya}. We formulate s_{xa} in a general way:

$$s_{xa} = s_{x,\ldots,i,\ldots,a} \tag{1}$$

For calculating the sign $s_{x,\ldots,i,\ldots,a}$,

$$s_{x,\ldots,i,\ldots,a} = s_{xx+1} \cdots s_{i-1i} \cdots s_{a-1a} \tag{2}$$

Similarly, s_{ya} can be written as:

$$s_{y,\ldots,j,\ldots,a} = s_{yy+1} \cdots s_{j-1j} \cdots s_{a-1a} \tag{3}$$

So, s_{yx} equals to:

$$s_{yx} = s_{x,\ldots,i,\ldots,a} \cdot s_{y,\ldots,j,\ldots,a} \tag{4}$$

2.4 Network Transforming

Now we transform the post signed network into user signed network. Given an edge $p_{ij} \in E_P$ for transform, there always exist three situations: ① $u_{p_i} = u_{p_j}$; ② $u_{p_i} \neq u_{p_j}\ and\ u_{ij} = 0$; ③$u_{p_i} \neq u_{p_j}\ and\ u_{ij} \neq 0$. For situation one, we just leave out the edge because u_{p_i} and u_{p_j} are the same user. For situation two, we set $u_{ij} = p_{ij}$. There is a little more complicated in situation three. If $|u_{ij}| = 1\ and\ |p_{ij}| < 1$ we just leave out the edge in that "1" which expresses explicit links should be more reliable than implicit links. Otherwise we set $u_{ij} = p_{ij}$ because the later commonly represents the relationship new status.

3 Identifying Opinion Leaders

Intuitively, the opinion leaders from online comments can be interpreted similar to the “authority” of a web page: A user has high influence if the sum of influence of his/her comments is high. Here, based on the PageTrust method [12], we propose a TrustRank method for handling positive and negative weight links.

The TrustRank is grounded on an idea that the distrust(negative) links may strengthen the possibility of leaving the network. In a $n \times n$ distrust matrix *P*, P_{ik} denotes the proportion of walkers in node *i* who distrusts node *k*. And the diagonal of P gives the proportion of walkers that distrust the node they are in. In that manner, (1-P_{ii}) represents the proportion of remaining walkers in node $i \in n$. Accompanied by the distrust matrix, the iteration process is defined as:

$$x_i^{(t+1)} = (1 - P_{ii}) \cdot \left[\alpha \sum_{j,(j,i) \in \ell^+} x_j^{(t)} \cdot w_{ji} / D_j + (1-\alpha) z_i \right], \tag{5}$$

where $\left[\alpha \sum_{j,(j,i) \in \ell^+} x_j^{(t)} / d_j + (1-\alpha) z_i \right]$ represents the traditional PageRank process and (1-P_{ii}) denotes the possibility of remaining the graph. The dynamic iteration of distrust matrix *P* is updated according to the equation:

$$\tilde{P}^{(t+1)} = T^{(t)} \cdot P^{(t)} \quad , \tag{6}$$

where *T* is the transition matrix and $T_{ij}^{(t)}$ is the ratio of node *i* who was *j* at time *t*,

$$T_{ij}^{(t)} = \frac{\alpha A_{ji}^{+} x_j^{(t)} \cdot w_{ji} / D_j + (1-\alpha) z_i x_j^{(t)}}{\alpha \sum_{k,(k,i) \in \ell^+} x_k^{(t)} \cdot w_{ki} / D_k + (1-\alpha) z_i} \tag{7}$$

In a signed network, there are three types of propagation information: positive-positive, negative-positive or reverse, negative-negative. The PageTrust method has considered the former two but ignores the negative-negative information. Here we introduce a new matrix to model the negative-negative propagation information. There are four types of atomic propagations in a network: direct propagation, co-citation, transpose trust and trust coupling [9]. Let *M* denote a connection matrix in a signed network with *n* nodes, the corresponding operators of four atomic propagations are: $M, M^T M, M^T, MM^T$. Let $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ be a vector weight for the four atomic propagations. Then four atomic propagations can be combined into a single matrix $C_{M,\beta}$:

$$C_{M,\alpha} = \beta_1 M + \beta_2 M^T M + \beta_3 M^T + \beta_4 MM^T, \tag{8}$$

where $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$. The *kth* iterative propagation can be denoted as $C_{M,\alpha}^k$. By introducing diverse damping factors (γ_t, γ_d) for trust and distrust matrix propagation, the detail matrix iterative propagation process can be defined as:

$$C_{M,\beta(i,j)}^k = \sum_{s=1}^{n} \gamma C_{M,\beta(i,s)}^{k-1} \cdot \gamma C_{M,\beta(s,j)} \tag{9}$$

And the γ can be formulated as:

$$\gamma = \begin{cases} \gamma_t & C^k_{M,\beta(i,j)} > 0 \\ \gamma_d & C^k_{M,\beta(i,j)} < 0 \end{cases} \tag{10}$$

Based on formulae (9) and (10), the negative-negative propagation information can be calculated by:

$$F_{(i,j)} = \sum_{k=1}^{m} \sum_{s} \gamma_d C^{k-1}_{M,\beta(i,s)} \cdot \gamma_d C_{M,\beta(s,j)} \text{ ,} \tag{11}$$

where m denotes the iteration depth and its value depends on actual situations. Then we combine the matrix F with the original matrix M to construct a new information matrix M' which has obtained the negative-negative information:

$$P^{(0)}_{(i,j)} = \begin{cases} M_{(i,j)} & \text{if } M_{(i,j)} \neq 0, \\ F_{(i,j)} & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{12}$$

4 Experiments

4.1 Datasets

The datasets are collected from the category of "Online Military Review" of the ChinaNet Military Forum[1] which is the largest and also the most active military forum in China. The forum provides a vote button for forum visitors to share agreements. We use the agreements for the gold opinion leaders. Here, we randomly downloaded about 1000 threads on 7 May, 2013. Then removing those comment posts which is less than two pages and get 53 threads. We extract some useful information, including user ID, post content, post floors and post votes (Since the crawler algorithm failed to download the vote information, we manually record the top 10 opinion leaders who have the most votes).

Table 1. Comparisons of top 10 opinion leaders between four models in thread 4

	UserID(votes)	Sim-PR	Vir-PR	PT	TR
1	Zjs16(2221)	1	2	8	**3**
2	Xysgy(1380)	6	10	11	**9**
3	Sfpy(867)	4	4	9	**1**
4	Fs_KK(562)	117	109	3	**14**
5	Afhdhg(277)	2	5	13	**4**
6	Yzqf618(173)	20	19	12	**7**
7	Dh_wgd(169)	21	20	14	**8**
8	Lsw(162)	7	6	17	**5**
9	Jlh(151)	48	44	7	**13**
10	Kw(151)	32	30	1	**12**

[1] http://club.china.com/data/threads/12171906/index.html

4.2 Results

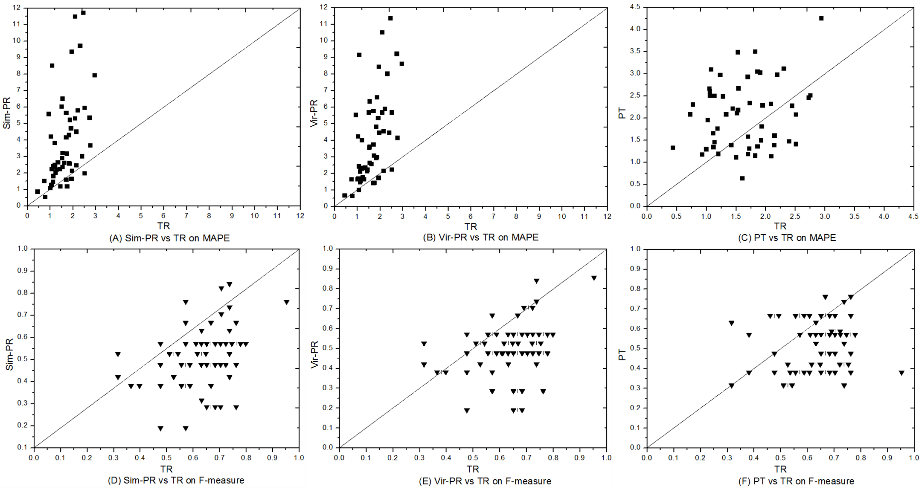
We compare the TrustRank (TR) model with three models for handling positive and negative links, including the Sim-PR, Vir-PR and PT. In addition, to give a clear comparison of the ranking result, we adopt Mean Absolute Percentage Error(MAPE) and F-measure. The MAPE is a common method for evaluating the difference between actual values and predicting values:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right|, \tag{13}$$

The lower the MAPE value, the better the ranking. The F-measure is a well-known evaluation method in information retrieval. The higher the F-measure, the better the result. Given the precision P and recall R, the F-measure is defined as:

$$F = \frac{2PR}{P+R} \tag{14}$$

For illustrating the ranking result of four models, we present the top 10 opinion leaders of the thread 4 as shown in table 1. The second column is the real top 10 opinion leaders and the last four columns are the ranking order of them in four models. The table illustrates that the TR model has a better ranking than the other three models. Specifically, 7 out of 10 nodes have a better order than the PT model; 5 out of 10 nodes obviously outperform both Sim-PR and Vir-PR models, while the other 5 nodes have an imminent ranking.



(A) Sim-PR vs TR on MAPE

(B) Vir-PR vs TR on MAPE

(C) PT vs TR on MAPE

(D) Sim-PR vs TR on F-measure

(E) Vir-PR vs TR on F-measure

(F) PT vs TR on F-measure

Fig. 3. Comparisons between TR and other three models in forum datasets. The top three figures show the results on MAPE; the bottom three figures show the results on F-measure.

Applying MAPE and F-measure for measuring the top 10 opinion leaders of 53 threads, figure 3 illustrates the detail results of four models. The figure shows that the TR model outperforms the other three models in two evaluation methods.

Specifically, 26 out of 53 threads' TR ranking is optimal in four models using MAPE and F-measure. 43 out of 53 threads' TR ranking is optimal in four models using MAPE or F-measure. For the remaining 10 threads, although they fail to obtain an optimal solution in TR model, they all get a suboptimal value using MAPE or F-measure. This lie in two reasons: 1) The TR model formulates negative links as negative influence, which degrades nodes that accept a large number of negative links the possibility of being important nodes. This satisfies the common sense that one opposed by a majority of people is less likely to being opinion leaders. So, the TR model outperforms both Sim-PR and Vir-PR models which treats negative links as none or positive influence. 2) The TR model also treats negative links as weak-positive influence, which can make up for mis-degrading nodes that accept negative links occasionally. So, the TR model is superior to PT model which only takes negative influence into consideration.

5 Conclusions

In this paper, we propose a novel method to identify opinion leaders from online comments based on both positive and negative opinions. The effectiveness of this method is validated on the online comments of a real forum.

Acknowledgements. This paper is supported in part by grants: NSFCs (National Natural Science Foundation of China) (61173075, 61272383 and 61370165), Natural Science Foundation of Guang Dong Province (S2013010014475), Strategic Emerging Industry Development Special Funds of Shenzhen (ZDSY20120613125401420), Shenzhen International Cooperation Research Funding (GJHZ20120613110641217) and Key Basic Research Foundation of Shenzhen (JC201005260118A).

References

[1] Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: Structure and algorithms. In: Proceedings of the 16th International Conference on World Wide Web, pp. 221–230 (2007)

[2] Ghosh, R., Lerman, K.: Predicting influential users in online social networks. Eprint arXiv: cs/1005.4882

[3] Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, 107–117 (1998)

[4] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) 46, 604–632 (1999)

[5] Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 261–270 (2010)

[6] Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R., Benevenuto, F.: Finding trendsetters in information networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1014–1022 (2012)

[7] Richardson, M., Agrawal, R., Domingos, P.: Trust management for the semantic web. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 351–368. Springer, Heidelberg (2003)

[8] Tai, A., Ching, W., Cheung, W.: On Computing Prestige in a Network with Negative Relations. International Journal of Applied Mathematical Sciences 2, 56–64 (2005)

[9] Guha, R., Kumar, R., Raghavan, P., Tomkins, A.: Propagation of trust and distrust. In: Proceedings of the 13th International Conference on World Wide Web, pp. 403–412 (2004)

[10] Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1606–1611 (2007)

[11] Heider, F.: Attitudes and cognitive organization. The Journal of Psychology 21, 107–112 (1946)

[12] De Kerchove, C., Dooren, P.: The PageTrust algorithm: how to rank web pages when negative links are allowed. In: SIAM: Data Mining Proceedings (2008)