OOLAM: An Opinion Oriented Link Analysis Model for Influence Persona Discovery

Keke Cai[1], Shenghua Bao[1], Zi Yang[2], Jie Tang[2], Rui Ma[1], Li Zhang[1], Zhong Su[1]
[1]IBM Research - China
[2]Department of Computer Science and Technology, Tsinghua University, China
[1]{caikeke, baoshhua, maruicrl, lizhang, suzhong}@cn.ibm.com
[2]yangzi@keg.cs.tsinghua.edu.cn, jietang@tsinghua.edu.cn

ABSTRACT

Social influence is a complex and subtle force that governs the dynamics of social networks. In the past years, a lot of research work has been conducted to understand the spread patterns of social influence. However, most of approaches assume that influence exists between users with active social interactions, but ignore the question of what kind of influence happens between them. As such one interesting and also fundamental question is raised here: "in a social network, could the social connection reflect users' influence from both positive and negative aspects?". To this end, an Opinion Oriented Link Analysis Model (OOLAM) is proposed in this paper to characterize users' influence personae in order to exhibit their distinguishing influence ability in the social network. In particular, three types of influence personae are generalized and the problem of influence persona discovery is formally defined. Within the OOLAM model, two factors, i.e., *opinion consistency* and *opinion creditability*, are defined to capture the persona information from public opinion perspective. Extensive experimental studies have been performed to demonstrate the effectiveness of the proposed approach on influence persona analysis using real web data sets.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-information filtering; H.3.4 [**Information Storage and Retrieval**]: Systems and Software-Information networks

General Terms

Algorithms, Measurement, Experimentation, Performance.

Keywords

Influence Persona Discovery, Link Analysis, Opinion Consistency, Creditability Analysis.

WSDM'11, February 9–12, 2011, Hong Kong, China.

1. INTRODUCTION

The social influence analysis is to study people's influence by means of analyzing the social interactions between people. It has attracted tremendous interest from both the sociology and data mining research communities. For example, Domingos and Richardson [9] proposed the influence maximization problem, in which the goal is to find a few "influential" members of the network. Kempe *et al.* [18] formalized the problem in discrete optimization and proposed three cascade models for influence propagation. However, most of approaches assume that influence exists between users with active social interactions, but ignore the question of what kind of influence happens between them. One interesting and also fundamental question has recently begun to be widely investigated: "in a social network, could the social interaction be described from both positive and negative aspects?" The answer is affirmative. As stated in Leskovec *et al.* [25] recent studies on signed networks [1], arbitrary social interaction on the Web involves both positive and negative relationships, which can affect the structure of on-line social networks.

The signed interactions between social individuals make it necessary to reconsider the influence analysis problem, since the vast majority of social influence researches are based on positive assumption and the generalized influence is therefore positive oriented by default. The signed social networks decide the diversity of the social influence. It thus becomes an interesting problem of how to explore the diverse characteristics of a user's social influence from the signed social connections.

In this paper, we aim to develop a better understanding of social influence when the social relationships between people can be positive and negative. Inspired by the idea of persona in sociology studies [13, 20], which show that the effects of the social influence by users with different personae are different, we give a definition of *social influence persona* to describe the social role or character of a user played in a social network. More specifically, three kinds of influence personae, which occur widely in social network, are generalized from public opinion perspective, including *Positive Persona*, *Negative Persona* and *Controversy Persona*. *Positive Persona* describes one kind of users with high positive influence, where their social links from others always indicate the friendship, support or approval. Comparatively speaking, *Negative Persona* represents users with high neg-

[1]Signed network is an important kind of network, links of which can reflect positive or negative social interactions.

ative influence, where their social links from others always express hostility, disagreement or distrust. The last kind of *Controversy Persona* represents a group of people who are liable to be challenged or supported by many. Generally, these three kinds of influence personae have similar features, and also different ones simultaneously. The similarity is in that they all represent people with huge influence, while the difference is in that their influence is reflected from different perspectives.

The problem in this work is referred to as *influence persona discovery* and two challenges are correspondingly identified. The first challenge is how to capture the influence from the interactions on signed social networks in a principled model. In this paper, this problem is explored from two aspects, namely "who are involved in the interactions" and "what have they said during the interaction". These two factors are derived from the common sense that people's influence can be convinced by the approvals gained from other reliable users. The concepts of *opinion consistency* and *opinion creditability* are therefore proposed. *opinion consistency* is to evaluate whether the social links from others are positive or not and *opinion creditability* is to measure to what degree the social link from others can be trusted.

Another challenge is how to leverage these two pieces of opinion information for influence persona analysis. In this research, an Opinion Oriented Link Analysis Model (OOLAM) is proposed to quantitatively estimate the influence persona in the social network. Specifically, we construct a bipartite graph, of which two sets of nodes respectively represent different set of social units with positive and negative interactions. An iterative process is performed on the bipartite graph to calculate the social influence from two complementary directions until a stationary distribution is found. The OOLAM model finally generates two ranking lists to describe people's positive and negative influence, based on which different influence personae can be easily recognized. For example, the positive persona refers to those who have higher positive ranking but lower negative ranking.

To validate the proposed OOLAM model, a series of experiments have been conducted on four public datasets from two popular online social media sites, i.e., Epinions [2] and Slashdot [3], where the obvious link polarity are provided to describe the opinions of online users for each other. For quantitative evaluations, we borrow the experimental framework articulated by Leskovec *et al.* [24] for edge sign prediction, wherein the identified social influence persona is considered as additional features to describe the signed social link. Experiment results strongly verify the effectiveness of the proposed model and at the same time arise some interesting questions deserved to be further studied.

The rest of the paper is structured as follows. Section 2 introduces the state-of-the-art of signed network and influence analysis. Section 3 gives the two hypotheses for our influence analysis in the context of social network. The proposed OOLAM model is then elaborated in Section 4 and distributed OOLAM algorithm is illustrated and discussed in Section 5. Section 6 demonstrates the experimental results. Finally Section 7 concludes this work with pointing out possible future work.

[2]http://www.epinions.com
[3]http://slashdot.org

2. RELATED WORK

2.1 Signed Links in Social Network

Traditional studies in social network take the assumption that the interactions between social users are positive. However, the negative relationships inherently exist in real online networks, e.g. the negative vote on Wikipedia [5], the distrust feedback on Epinion [15], the "foes" declaration on Slashdot [4, 22, 23], and other implicit disagreement hidden behind online discussion [30]. Recently, a lot of studies are involved in investigating both the positive and negative relationships in social context. Leskovec *et al.* [24, 25, 26] implemented a series exploration on this topic. They found that the interplay between positive and negative relationships indeed affects the structure of on-line social networks and then defined a new theory of status to explain the observed edge signs [25]. Another problem they investigated is how to realize the prediction of positive and negative link given the underlying social networks [24]. A machine learning approach is then implemented based on a collection of features observed from link structure and social psychology. Controversial study is another research topic on soical issues. For example, works of Massa and Avesani [27] studied the global and local trust metrics for controversial user and suggested to use Local Trust Metrics to access the trustworthiness of a user.

2.2 Social Influence Analysis

Identifying influential users in the interactive network has been highlighted as a key issue for study. Graph-theoretic approaches have been widely applied to solve this problem, where influence analysis is transferred to a link analysis problem and the directed links are used to represent the influences between each other.

Heuristics approaches that focus on exploring the link structure are extensively used for influence analysis. Network centrality measure is one of the most representative mechanisms. Network centrality [6] focused on node structure of a network and decided node importance based on its structure location. Different centrality measures have been applied, including degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality [10] etc.

PageRank [29] and HIT [19] etc link analysis algorithms are also adopted to solve influence or expertise ranking problem. PageRank algorithm measures the influence by analyzing the underlying hyperlink structure of entities and the HIT algorithm discovers authoritative and hub entities based on a mutual reinforcement principle. However, since PageRank and HIT algorithms are traditional webpage adaptive approaches, as mentioned in previous work [1], they do not work well for sparsely linked structures. Thus, variations of PageRank and HITS are proposed in succession. Adding implicit links [21] to increase the link density is a widely applied solution.

In recent years, finding the maximization of information diffusion in the network is explored to find the most influential entities [14]. Two basic diffusion models have been widely applied, i.e., linear threshold model [14, 3] and cascade model [11, 12]. These works studied the diffusion problem from different aspects and came to different conclusions. Java *et al.* [16] found that PageRank can be considered as an inexpensive approximation to the greedy heuristic in selecting the initial target set for activation. In the study of [26],

a new scheme called "Cost-Effective Lazy Forward" (CELF) is proposed for new seeds selection. Greedy algorithms [18, 7] applied for influence analysis proved that they are outperforming the classic degree and centrality-based heuristics in influence spread. Influence propagation method also has been studied in topic-level influence analysis. In the study of Tang *et al.* [31], the Topical Affinity Propagation (TAP) is proposed to model the topic-level social influence on large networks.

Link-based and probabilistic-based approaches have been shown to be successful in ranking entity influence on the graph, however, most of them focus on the structure of graph and hardly consider the intrinsic properties hidden behind the entities. In the study of Agarwal *et al.* [1], except the link structure, several post level statistics, like blog post recognition, novelty, eloquence and comments etc information, are also explored and experiments proved that such object-oriented statistic analysis is quite useful for influence analysis. The approach disscussed in our paper,

Approaches discussed above explored influence problem from various aspects, while the signed interactions between social individuals has been largely unexplored. This paper aims to give some initial explorations on influence analysis with consideration of the signed links in social network. There are crucial differences compared with previous approaches. Firstly, most previous graph-based influence analysis algorithms assume that the presence of a link from u to v is an evidence that u is influenced by v [16]. However, it is not always a fact. Controversy in user discussion make it possible that link from u to v actually represents u disagrees with v. Secondly, traditional influence analysis approaches assume that the link from u to v is reliable. However, in reality, the links from an unreliable user, e.g., a spammer, are most likely unreliable. Motivated by the observations above, we propose to incorporate opinion concept into influence analysis to reveal the intrinsic features of user influence in social connection.

3. USER INFLUENCE THROUGH SOCIAL INTERACTION

Before the introduction of detailed algorithm, we firstly give some notations involved in our social interaction and influence analysis.

3.1 Social Interaction

Basically, online social interaction can be explicit, such as the "friends" or "foes" relationship between users on Slashdot, or implicit, such as the reply relationship among users in online forums etc. But whatever kind of interaction, two kinds of elements are always involved, namely user and opinion. Figure 1 gives an example to illustrate the elements involved in the social interaction. In our following discussion, if not stated otherwise, we treat social interaction as directed.

As shown in Figure 1, a *response* that is presented as the link between users represents a social action from one user to another. Furthermore, a response is said to be a *positive response* if it supports the opinion of the other one, else is said to be a *negative response*. A user is therefore said to have *positive influence* if he receives positive response, else is said to have *negative influence*. It can be seen from Figure 1 that a user can simultaneously have positive and negative influence.

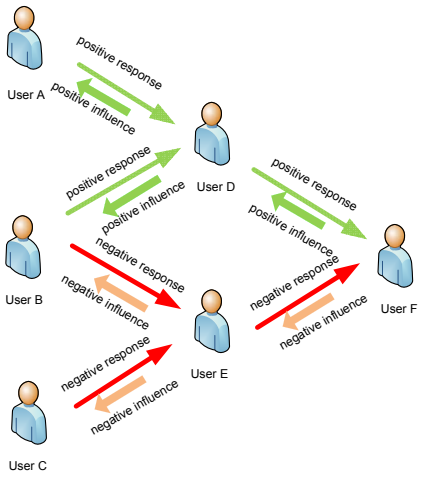


Figure 1: Elements involved in user social interaction.

Based on the above definitions, the interactions between users can be represented by a user graph $G = \{A, E\}$, where A is the union of all users and (a_i, a_j) is a directed arc in E if user $a_i \in A$ has ever generated one or more social responses to user $a_j \in A$.

3.2 User Influence and Opinion Consistency

Intuitively, the simplest way to measure the influence of a user is to count the number of responses he/she gets from others. It is perceived that the given remark exert much influence on social community if it receives much social attentions. However, such an approach does not consider a fact that not all responses are positive, so that users receiving many negative responses can be incorrectly regarded to have the same influence capability with those receiving many positive responses.

As defined in Wikipedia, "Social influence occurs when an individual's thoughts or actions are affected by other people"[4]. We then believe that a user with strong positive influence should be someone who can always induce others to behave in a similar way, and vice versa. Figure 2 gives an example to explain the effect of opinion consistency on user influence analysis. As shown in Figure 2, although user A and B similarly get two responses from others, the influence of A is considered more positive since it gains more positive responses than B.

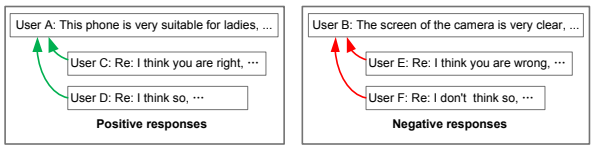


Figure 2: Opinion consistency in user discussions.

Opinion consistency can be captured from online social

[4]http://en.wikipedia.org/wiki/Social_influence

interactions directly or indirectly. For some social networks with explicit link signs, such as Slashdot etc, we can easily capture it through the relation tags manually labeled by users. However, for another kind of online resources, like online forums etc, more action is needed to infer the hidden opinion information from user discussion content. It can be imagined that user's online discussions indeed are some kind of opinion exchange, from which the consistency between user's opinions can be detected. In this paper, two approaches are discussed for opinion consistency detection from text analysis perspective. But, since this task is not the focus of our paper, we just briefly present the basic idea. Further studies will be discussed in future works.

Taking the forum discussion as an example, where user's social interaction is concretely manifested by the interactive messages, the approaches for opinion consistency detection are implemented as follows.

1. *Sentiment based consistency detection.* Given a message M_A that replies to message M_B, sentiment based consistency detection will consider M_A holding the same opinion with M_B if and only if they shall the same sentiment about the discussed object. The accumulated consistency score from interactive messages is then used to describe the opinion consistency between users. Sentiment based consistency detection is easy to be implemented, but its performance is limited by the sentiment analysis and opinion mining technique [30].

2. *Rule based consistency detection.* Observations from online discussions show that some agree and disagree indicators contained in messages can directly indicate the consistency (or not) between messages. Table 1 lists some examples of agree and disagree indicators captured from online forum. By referring to these explicit indicators, the rule based approach can be implemented for consistency detection. This approach obviously can accurately capture user's opinion expressed through messages. However, compared with sentiment based consistency detection, rule based approach is lack of flexibility and scalability. So, it raises an interesting question of whether the rule-based approach can be used to improve traditional learning-based approach for sentiment analysis. If the answer is affirmative, the applicability of sentiment based consistency detection can be improved.

Table 1: Examples of agree and disagree indicators

Agree indicator	right out of my mouth sound promising likely correct well said could not agree with you more ...
Disagree indicator	without merit bad idea obviously untrue you are wrong not necessarily true ...

3.3 User Influence and Opinion Creditability

Opinion consistency can differentiate social responses from positive and negative perspectives, but it treats all responses from different users equally. Commonly, the statements from an influential expert are more reliable than those from an indifferent person. Therefore, to decide the effect of responses on influence decision, opinion creditability is another important factor should be considered.

Opinion creditability refers to the idea that user's influence depends largely on the responsive users with high influence. Figure 3 illustrates this idea. From this figure, it can be seen that the support from reliable user E to A is more convincible than that from unreliable user F to C. Influence of A rather than B is then proved. Similarly, the negative response from E to B convincingly negates B's influence, but it is not the same case for D because of the low creditability of F.

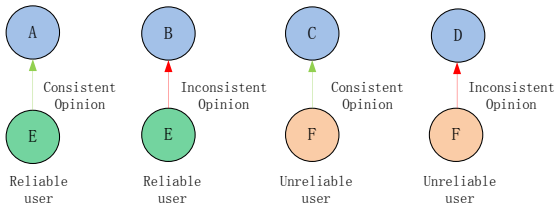


Figure 3: Effect of opinion creditability on influence analysis.

4. OOLAM MODEL FOR INFLUENCE PERSONA DISCOVERY

After the discussion of two elements involved in our influence analysis, this section gives a detailed description of the opinion oriented link analysis model, followed by influence persona discovery.

4.1 Positive/Negative Reinforcement

In this paper, a bi-partite graph model is proposed for user influence analysis, in which the social interactions among users will be observed from opinion consistency perspective and two kinds of reinforcement, namely positive and negative reinforcement, are studied on the basis of opinion creditability.

Based on the assumption of opinion consistency, we divide user graph $G = \{A, E\}$ into two weighted graphs, $G^P = \{A_P, E_P\}$ and $G^N = \{A_N, E_N\}$, where $A_P, A_N \subseteq A$ and $A_P \cap A_N \neq null$, (a_j, a_i) is a directed arc in E_P if user $a_j \in A_P$ has responded to user $a_i \in A_P$ with positive opinion, similarly (a_j, a_i) is a directed arc in E_N if user $a_j \in A_N$ has responded to user $a_i \in A_N$ with negative opinion. The weights $w_P^A(a_j, a_i)$ and $w_N^A(a_j, a_i)$ associated with each link in G^P and G^N describe the response intensity from a_j to a_i. Generally, graph G^P and G^N reflect the positive and negative interaction among users respectively, based on which users will be assigned with two influence scores to describe their positive and negative characteristics on influence. These two conflicting scores are viewed as the determinative factors to define user's creditability. Basically, users with high positive influence and low negative influence are considered as creditable ones.

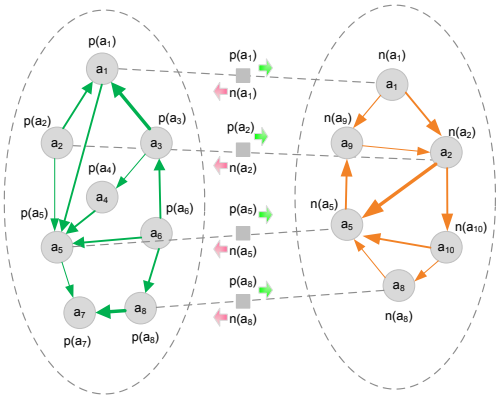


Figure 4: OOLAM model.

Figure 4 gives an example to illustrate the proposed model for influence analysis. The observed objects consist of 10 users $\{a_1, a_2, \ldots, a_{10}\}$ and are divided into G^P and G^N, where $G^P = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$ contains all users with positive interactions and $G^N = \{a_1, a_2, a_5, a_8, a_9, a_{10}\}$ contains all users with negative interactions. The weight of the link represents the connectivity strength between users. $p(a_i)$ and $n(a_i)$ respectively represent the positive and negative influence score of user a_i, which is the basis of the communication between G^P and G^N. Table 2 gives an example to explain the factors used for calculating $p(a_5)$ and $n(a_5)$. The column of "Weight of in-link" describes the positive and negative responses a_5 received from others and the column of "Opinion Creditability" denotes the creditability of each responsive user which is computed as the subtraction of their positive and negative influence score. Specifically, for users only contained in graph G^N, a small value, such as the parameter ϵ in Table 2, will be assigned to describe its opinion creditability. The purpose is to limit its opinion effect on others.

Table 2: Examples of the influence factors in Bi-OOIA model

	Weight of in-link	Opinion creditability
$p(a_5)$	$w_P^A(a_1, a_5)$	$a_1 : p(a_1) - n(a_1)$
	$w_P^A(a_2, a_5)$	$a_2 : p(a_2) - n(a_2)$
	$w_P^A(a_4, a_5)$	$a_4 : p(a_4)$
	$w_P^A(a_6, a_5)$	$a_6 : p(a_6) - n(a_6)$
$n(a_5)$	$w_N^A(a_2, a_5)$	$a_2 : p(a_2) - n(a_2)$
	$w_N^A(a_8, a_5)$	$a_8 : p(a_8) - n(a_8)$
	$w_N^A(a_{10}, a_5)$	$a_{10} : \epsilon$

4.2 OOLAM Algorithm

Let $I_{A_P} = (p(a_1), p(a_2), \ldots, p(a_t))$ and $I_{A_N} = (n(a_1), n(a_2), \ldots, n(a_u))$ be the positive and negative ranking of nodes in G^P and G^N, where $t = |I_{A_P}|$ and $u = |I_{A_N}|$, the simplest approach to evaluate $p(a_i)$ and $n(a_i)$ is to see how many positive and negative responses a_i get from its neighbors in G^P and G^N.

$$p(a_i) = \sum_{(a_j, a_i) \in E_P} w_P^A(a_j, a_i) \tag{1}$$

$$n(a_i) = \sum_{(a_j, a_i) \in E_P} w_N^A(a_j, a_i) \tag{2}$$

To implement the idea of opinion creditability, we need to consider the creditability of the responsive user into Eq. 1 and Eq. 2. Basically, user's opinion creditability $R(a_i)$ is decided by its positive and negative influence score, namely $p(a_i)$ and $n(a_i)$, and can be defined as:

$$R(a_i) = \max(\epsilon, p(a_i) - n(a_i)) \tag{3}$$

We define M_{A_P} and M_{A_N} as the adjacency matrix to describe user relationship in graph G^P and G^N:

$$M_{A_P}(i, j) = w_P^A(a_i, a_j) \tag{4}$$

$$M_{A_N}(i, j) = w_N^A(a_i, a_j) \tag{5}$$

Then the user influence scores can be calculated iteratively as below:

$$I_{A_P}^{k+1} = R^k * M_{A_P} \tag{6}$$

$$I_{A_N}^{k+1} = R^k * M_{A_N} \tag{7}$$

Further considering the damping factor d into ranking function, the equation is as follows:

$$I_{A_P}^{k+1} = \frac{1-d}{t} + d * R^k * M_{A_P} \tag{8}$$

$$I_{A_N}^{k+1} = \frac{1-d}{u} + d * R^k * M_{A_N} \tag{9}$$

The OOLAM algorithm is guaranteed to converge. For the ease of description, we simply assume $R(a_i) = p(a_i) - n(a_i)$. Then the convergence of OOLAM can be proofed as below: By subtracting Eq. 9 from Eq. 8, we have:

$$I_{A_P}^{k+1} - I_{A_N}^{k+1} =$$
$$(\frac{1-d}{u} + d * R^k * M_{A_P}) - (\frac{1-d}{t} + d * R^k * M_{A_N})$$
$$\Rightarrow R^{k+1} = (1-d)(\frac{1}{u} - \frac{1}{t}) + dR^k(M_{A_P} - M_{A_N}) \tag{10}$$

Letting $C = \frac{1}{u} - \frac{1}{t}$ and $M_{A_{PN}} = M_{A_P} - M_{A_N}$, the Eq. 10 can be rewritten as:

$$R^{k+1} = (1-d)C + dR^k M_{A_{PN}} \tag{11}$$

Then the iteration of opinion creditability R has a similar form with the traditional PageRank [29] algorithm whose convergence has been proven before. As R^k converges, we can easily derive that $I_{A_P}^{k+1}$ and $I_{A_N}^{k+1}$, which both depend on variable R^k only, will also converge. Note that PageRank acceleration approaches can also be applied here [28].

In summary, the final OOLAM algorithm is shown as Algorithm 1.

4.3 User Influence Persona

The OOLAM model generates two ranking lists to describe user's influence characteristics from positive and negative perspectively, based on which three types of persona can be defined.

Algorithm 1: Bi-partite Opinion Oriented Link Analysis

input : Social interactions between users
output : positive and negative influence ranking lists

1 Construct social network graph G based on user social interactions;
2 Divide G into two sub-graphs G^P and G^N according to the opinion consistency between users;
3 Let R be the opinion reliability of each node in G, I_{A_P} and I_{A_N} be the ranking of nodes in positive and negative graph respectively;
4 **foreach** $k \leftarrow 1$ *to* n **do**
5 $\quad I_{A_P}^{k+1} = \frac{1-d}{t} + d * R^k * M_{A_P}$
6 $\quad I_{A_N}^{k+1} = \frac{1-d}{u} + d * R^k * M_{A_N}$
7 Return I_{A_P} and I_{A_N}

1. *Positive Persona*. The first type of persona represents users whose opinions are always accepted by others. Having relatively high positive influence and low negative influence is the characteristic of this type of persona.

2. *Controversy Persona*. The second type of persona describes users who always raise controversial issues. The most characteristic of this type of persona is that their positive influence and negative influence are finely balanced.

3. *Negative Persona*. Contrary to positive persona, the third type of negative user is with relatively low positive influence and high negative influence. Opinions of this type of users are always denied by others.

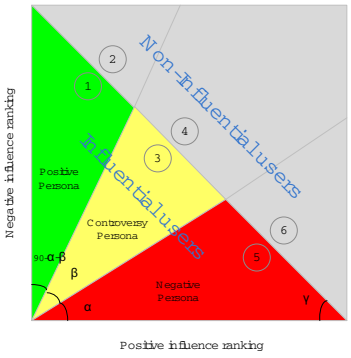


Figure 5: Area division for influence persona discovery.

In this paper, a two-dimensional coordinate system is efficiently used to depict these three types of personae. Based on the two generated influence ranking lists, each user can be mapped to a two-dimensional coordinate point, where the x-axis corresponds to user's positive influence ranking and y-axis corresponds to the negative influence ranking. After the coordinate mapping, all the coordinate points can be divided into six areas, as shown in 5. In this figure, points in green area 1 correspond to users with the first type of influence persona since their relative high positive influence and low negative influence. Similarly, points in yellow area 3 and red area 5 respectively represent users with most controversial and negative characteristic of influence. Obviously, compared with points in other 2, 4 and 6 areas, points area 1, 3 and 5 are deserved to be noticed because of their significant characteristic of influence, and therefore go by the general name of *influential users*, while points in Area 2, 4 and 6 are all treated as *non-influential users*.

Phase	Key	Value
Map Input	a_i	$\langle p^k(a_i), n^k(a_i)\rangle$
Map Output / Reduce Input	a_j	$\langle \frac{1-d}{t} + dR^k(a_i)w_P^A(a_i,a_j),$ $\frac{1-d}{u} + dR^k(a_i)w_N^A(a_i,a_j)\rangle$
Reduce Output	a_i	$\langle p^{k+1}(a_i), n^{k+1}(a_i)\rangle$

Table 3: Key value definitions for each computation phase

5. DISTRIBUTED OOLAM ALGORITHM

As a social network may contain millions of users and hundreds of millions of social ties between users, it is impractical to proceed a OOLAM algorithm from such a huge data using a single machine. To address this challenge, we deploy the learning task on a distributed system under the map-reduce programming model [8]. Map-Reduce is a programming model for distributed processing of large data sets. In the map stage, each machine (called a process node) receives a subset of data as input and produces a set of intermediate key/value pairs. In the reduce stage, each process node merges all intermediate values associated with the same intermediate key and outputs the final computation results. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key.

For distributing the OOLAM Algorithm, we first duplicate a complete graph G onto each single process node, and a portion of nodes are assigned to each node at each iteration randomly and automatically for calculating the reliabilities I_{A_P} and I_{A_N} corresponding to that portion of nodes. And then, the map stage and the reduce stage can be defined as follows.

In the map stage, each process node scans the reliabilities $p^k(a_i)$ and $n^k(a_i)$ of the k-th iteration. The map function is defined in Table 5. We note that for each input key/value pair it issues a series of intermediate key/value pairs for each a_j such that $(a_i, a_j) \in E_P \cap E_N$. If $(a_i, a_j) \notin E_P$ or $(a_i, a_j) \notin E_N$, the corresponding issued value takes 0.

In the reduce stage, each process node collects all values associated with an intermediate key to generate new pair by summing up all the intermediate values, which corresponds to the input in the map phase for the next iteration. Thus, the one time map-reduce process corresponds to one iteration in our Distributed OOLAM algorithm.

6. EXPERIMENTAL STUDY

To quantitatively measure the performance of the proposed approach for influence persona discovery, we borrow the experimental framework articulated by Leskovec *et al.*, in which the machine-learning approach is implmented for edge sign prediction by combining the evidence from various edge features. By treating influence persona as new features for classification, it is hoped to see that user associated influ-

ence persona can offer more evidence for social relationship prediction. In our experiments, the task of edge sign prediction is considered as a binary classification implemented by svmlight[5] [17] and classification accuracy is taken as the main measure for evaluation.

6.1 Experiment Setup

The experimental data are generated based on the public data released by Leskovec *et al.* [24], which include four data sets from two popular online social media sites, i.e., Epinions and Slashdot. These network data contain explicit signs of links to indicate the attitudes of online users for each other and is especially suitable for our experiments. Firstly, these network data reflect a sense of direct social relationship between users and secondly, the provided social relationships are assigned with explicit signs, the noise caused by opinion consistency detection therefore can be avoided. Table 4 shows the detailed statistics of these datasets. As stated in [24], the overwhelming majority of the edges in the selected experimental data are positive and the random guessing can achieve approximately 80% accuracy. To reduce the effect of randomness on classification, we follow the process of Leskovec *et al.* and create a *balanced dataset* with equal numbers of positive and negative edges for training and testing.

Table 4: Dataset statistics

Name	Node	Edges	Description
Epinions	131,828	841,372	Epinions signed social network
Slashdot 1	77,350	516,575	Slashdot Zoo signed social network from November 6, 2008
Slashdot 2	82,140	54,9202	Slashdot Zoo signed social network from February 21, 2009
Slashdot 3	81,867	54,5671	Slashdot Zoo signed social network from February 16, 2009

We implement the OOLAM algorithm and run on above datasets by setting damping factor d to 0.95. The algorithm converges after average 15 iterations. Based on the generated positive and negative ranking lists, we can easily assign each user with a kind of influence persona according to the rules introduced in Section 4.3. In the following analysis, we set k to the size of experimental data, and the angle of α, β, and γ to 30, 30 and 45, respectively. Table 5 gives the distribution of nodes in each area. From this table, it is easy to see that users in Area 2, 4 and 6 actually represent the majority of web users that are always paid little attention by others. In contrast, the proportion of influential users in Area 1, 3 and 5 is relatively small. The distribution of influential and non-influential users follows a power-law distribution, which indeed is a natural phenomenon in real situations.

6.2 Comparsion With Baseline Classifier

A baseline classifier is constructed by referring to the five basic degree features discussed in [24]. Specifically, given the edge E from node u to v, a 5-dimensional vector is constructed, where features of $d^+_{in}(v)$ and $d^-_{in}(v)$ denote the number of positive and negative incoming links to v, features of $d^+_{out}(u)$ and $d^-_{out}(u)$ represent the number of positive and negative outgoing links from u and feature $C(u,v)$ denotes the total number of common neighbors of u and v with no consideration of edge direction. Formally, E can be represented by the feature vector $\{d^+_{in}(v), d^-_{in}(v), d^+_{out}(u), d^-_{out}(u), C(u,v)\}$.

[5] `http://svmlight.joachims.org/`

Table 5: Distribution of users in each area

	Epinion	Slashdot1	Slashdot2	Slashdot3
Area1	3,290	3,280	3,230	3,263
Area2	52,810	26,735	29,535	29,390
Area3	5,159	7,448	7,143	7,137
Area4	47,227	21,750	23,308	23,211
Area5	8,641	9,582	10,201	10,154
Area6	14,701	8,555	8,723	8,712

Features involved in the baseline classifier aim to generalize the interaction characteristics between nodes. Comparatively, influence persona provides another kind of characteristics about nodes themselves. We therefore construct a compared classifier named IPClassifier, where the information about influence persona of the linked nodes is also leveraged to construct feature vector of edge. The basic idea is that a user with high positive influence is more likely to receive positive links from others. Then, given the edge E from node u to v, a 7-dimensional feature vector $\{d^+_{in}(v), d^-_{in}(v), d^+_{out}(u), d^-_{out}(u), C(u,v), I_p(v), I_p(u)\}$ is generated, where the new added features of $I_p(v)$ and $I_p(u)$ denote the influence persona of v and u respectively.

The average accuracy of classification throughout the four datasets is shown in Figure 6, where the results are compared between the baseline classifier and IPclassifier for different size of training and testing data. It can be seen from Figure 6 that IPClassifier significantly outperform the baseline classifier on all data sets. It is also noted that the performance of baseline classifier has no obvious change with the increase of training data, while the performance of IPclassifier has a dramatic improvement in the intial stage and gradually reaches a stable state. All these evidences demonstrate that the proposed approach for influence persona discovery is effective, which can efficiently describe the characteristic of online user from influence perspective.

6.3 Evaluation of Link Polarity in Influence Analysis

Earlier in the above discussion, influential analysis has been a key issue for study and a lot of graph-theoretic approaches have been proposed successively. One natural and interesting question is then raised "could the prediction performance be improved if the influence features are generalized from traditional influence model?". To get the answer, a compared classifier based on traditional influence analyzer called TRIClassifier is constructed. As we know, traditional influence model always generate single ranking list. To incorporate such kind of influence information for prediction, the ranking list is firstly segmented into several sections and nodes belonging to the same sections are assigned with an unique section ID. Then, a 7-dimensional vector is constructed and formed as $\{d^+_{in}(v), d^-_{in}(v), d^+_{out}(u), d^-_{out}(u), C(u,v), SID(v), SID(u)\}$, where $SID(u)$ and $SID(v)$ correspond to the section ID of nodes u and v. In our experi-

ment, we apply PageRank algorithm to generate the ranking list. Basically, approach that is capable of solving influence or expertise ranking problem can be applied here.

Figure 7 gives the comparisons between IPClassifier and TRIClassifier. We can see that the performance of TRIClassifier is largely lower than that of IPClassifier on most data sets. It reflects to some degree that traditional influence analysis approach cannot adapt very well to networks with significant signed edge. To this end, it is necessary to separately consider the positive and negative links between users for influence estimation.

6.4 Evaluation of Positive/Negative Reinforcement in Influence Analysis

Experiments above have proved the necessary of opinion polarity for influence analysis. As discussed above, another important factor needed to be considered in our analysis model, except the link polarity, is the interactions between the separated social graphs. Thus, another problem comes up "whether such kind of information is helpful in depicting user's influence?". In the following experiments, we will turn to this problem.

Given two graphs separated by link polarity, influence analysis with no consideration of mutual reinforcement is implemented as follows. Firstly, apply influence anlaysis on each graph independently, then, assign each graph node two independent scores, and finally calculate node influence persona by using the same approach defined in section 4.3. To distinguish with the influence persona defined in our approach, influence persona here is denoted as independent influence persona. By leveraging the independent influence persona into edge description, the third compared predictor named STRIClassifier is constructed. The feature vector of edge from u to v therefore can be described as $\{d^+_{in}(v),\ d^-_{in}(v),\ d^+_{out}(u),\ d^-_{out}(u),\ C(u,v),\ I_{idp}(u),\ I_{idp}(v)\}$, where $I_{idp}(u)$ and $I_{idp}(v)$ correspond to the independent influence persona of nodes u and v.

Figure 8 gives the classification accuracy of the compared approaches of IPClassifier and STRIClassifier. However, the result is not as expected. The peformance of IPClassifier and STRIClassifier are almost at the same level. It is an interesting phenomena deserved to be studied. Compared with STRIClassifier, IPClassifier will consider the mutual reinforcement between the separated graphs during influence estimation. The communication density between nodes therefore becomes very important. The statistics of nodes communication in graphs is shown in Figure 9, from which some observations are summarized as follows. First, there is only small part of nodes with both positive and negative in-link. It means that majority of nodes in the two graphs have only one influence score. Second, limited transmission happens between nodes. It can been seen that only small part of nodes which have positive in-link can transmit its positive influence to negative graph and similarly only small part of nodes which have negative in-link can transmit its negative influence to positive graph. As a result, it severely restrains the implementation of influence reinforcement. The sparse communication between nodes directly result in similar performance of IPClassifier and STRIClassifier. It therefore raises an interesting issue for our future work that how to enhance the adaptability of the influence model to various data structure.

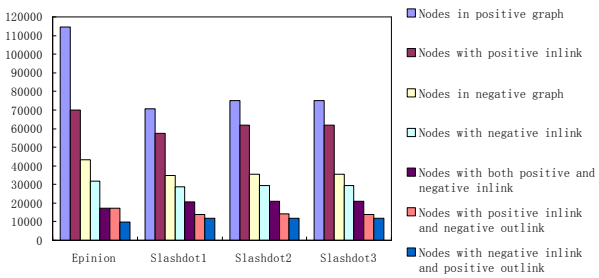
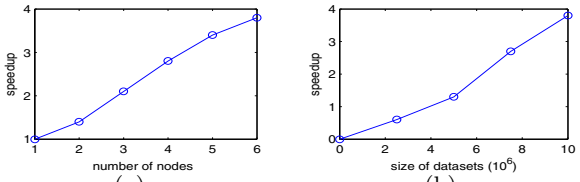


Figure 9: Nodes distribution in the positive and negative graph.

6.5 Scalability Performance

The distributed learning algorithm is also implemented under the Map-Reduce programming model using the Hadoop platform[6]. We perform the distributed train on 6 computer nodes (24 CPU cores) with AMD processors (2.3GHz) and 48GB memory in total.

We evaluate the speedup of the distributed learning algorithm on the cluster of 6 nodes using the complete data set with different sizes of nodes. Moreover, we perform the distributed learning algorithm on all the computational nodes, but with different size of data set. The results are shown in Figure 10(a) and Figure 10(b). It is shown that when the size of data set increases to nearly one million edges, the distributed learning starts to show a good parallel efficiency (speedup > 4). This confirms that the distributed OOLAM algorithm like many distributed learning algorithms is preferable to large-scale data sets.



(a) (b)

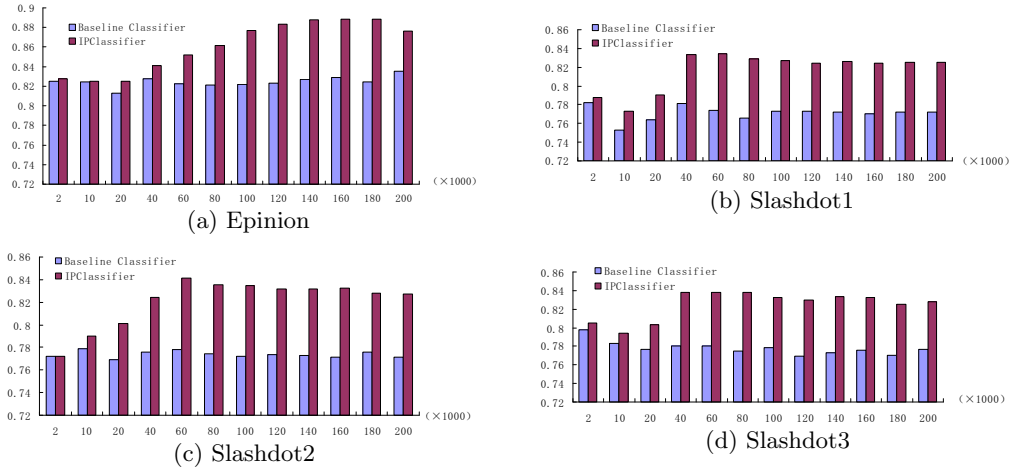
Figure 10: Scalability Performance. (a)Scalability performance using different number of nodes. (b)calability performance on data sets of different sizes

7. CONCLUSIONS AND FUTURE WORK

With the fast development of Web 2.0, more and more Web users share their opinions online through various social channels. Such information is quite valuable in reflecting the social activities and draws large attention from both industry and academia. In this paper, an opinion-oriented link analysis model is proposed for social influence analysis, wherein three kinds of influence persona are discovered. The contributions of this work can be summarized as follows.

1. The proposal of deriving user influence persona from the public opinion's perspective. Three kinds of persona are studied based on their influence effects, i.e., positive users, negative users and controversial users.

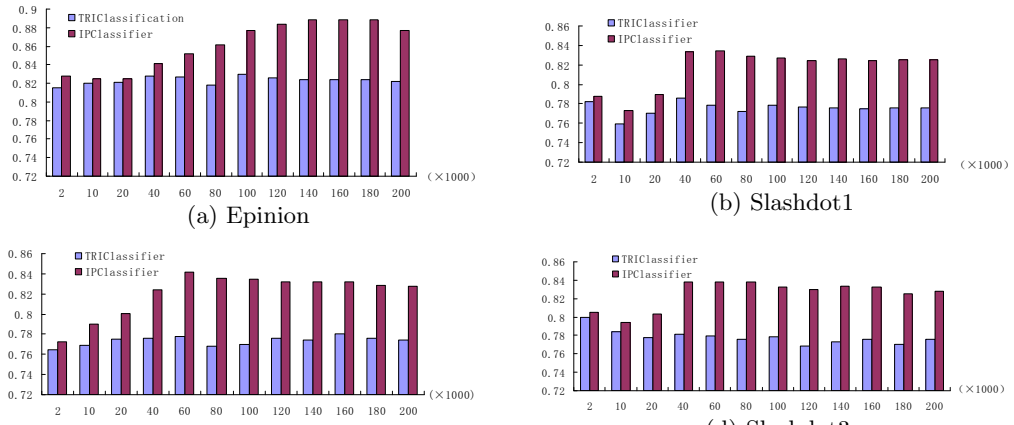
[6] `http://hadoop.apache.org/`



(a) Epinion (b) Slashdot1

(c) Slashdot2 (d) Slashdot3

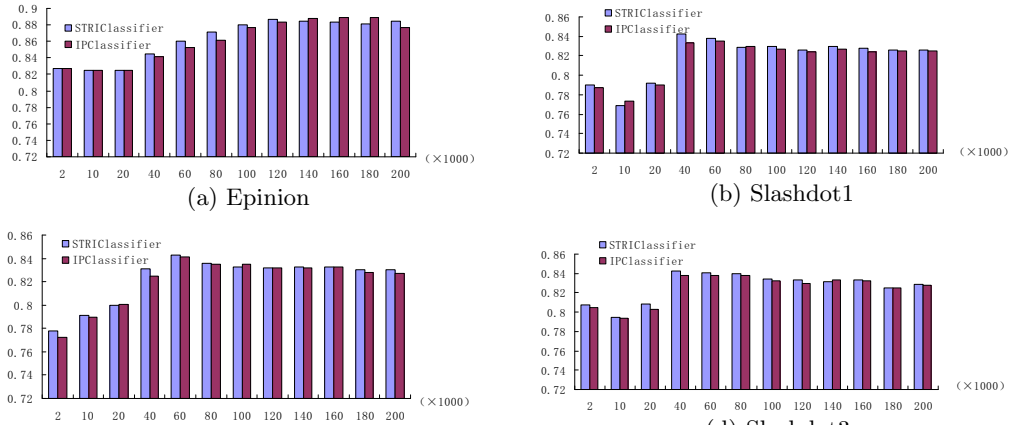
Figure 6: Classification Accuracy Compared between the baseline and IPClassifier.



(a) Epinion (b) Slashdot1

(c) Slashdot2 (d) Slashdot3

Figure 7: Classification Accuracy Compared between IPClassifier and TRIClassifier.



(a) Epinion (b) Slashdot1

(c) Slashdot2 (d) Slashdot3

Figure 8: Classification Accuracy Compared between IPClassifier and STRIClassifier.

2. The proposal of OOLAM algorithm for opinion-oriented link analysis and the mapping between influence ranking and influence persona. It helps capture the user influence more accurately by propagating both positive influence and negative influence iteratively.

3. The extensive experiment of influence persona discovery. Experiment results at one side strongly verify the effectiveness of the proposed model and at another side arise some interesting questions deserved to be further studied.

Generally speaking, it is a first try on influence persona discovery and there is still a long way to go. Model adaptability and inference drifting etc problems need to be studied more deeply in our future work. The discovered influence persona can benefit a number of applications, For example, it is much easier to get a whole picture of a forum by sampling the key messages published by influential users. The discovered persona can also be used to enhance existing probabilistic models for expert finding as a prior weighting [2].

8. REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *WSDM'08*, pages 207–218, 2008.

[2] S. Bao, H. Duan, Q. Zhou, M. Xiong, Y. Cao, and Y. Yu. A probabilistic model for fine-grained expert search. In *ACL'08*, pages 914–922, Columbus, Ohio, June 2008.

[3] E. Berger. Dynamic monopolies of constant size. *Journal of Combinatorial Theory*, 83(2):191–200, 2001.

[4] M. J. Brzozowski, T. Hogg, and G. Szabó. Friends and foes: ideological social networking. In *CHI'08*, pages 817–820, 2008.

[5] M. Burke and R. Kraut. Mopping up: modeling wikipedia promotion decisions. In *CSCW'08*, pages 27–36, New York, NY, USA, 2008. ACM.

[6] M. Chau and J. J. Xu. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Man-Machine Studies*, 65(1):57–70, 2007.

[7] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD'09*, pages 199–208, 2009.

[8] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51:107–113, January 2008.

[9] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD'01*, pages 57–66, 2001.

[10] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215– 239, 1979.

[11] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 3(12):211–223, 2001.

[12] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001.

[13] M. S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[14] D. Gruhl, D. Liben-Nowell, R. V. Guha, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explorations*, 6(2):43–52, 2004.

[15] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW'04*, pages 403–412, New York, NY, USA, 2004. ACM.

[16] A. Java, P. Kolari, T. Finin, and T. Oates. Modeling the spread of influence on the blogosphere. In *WWW 2006 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[17] T. Joachims. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms.* Kluwer/Springer, 2002.

[18] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD'03*, pages 137–146, 2003.

[19] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JOURNAL OF THE ACM*, 46(5):604–632, 1999.

[20] D. Krackhardt. *The Strength of Strong Ties: The Importance of Philos in Organizations*, pages 216–239. Harvard Business School Press, Boston, MA.

[21] A. Kritikopoulos, M. Sideri, and I. Varlamis. Blogrank: Ranking weblogs based on connectivity and similarity features. *CoRR*, abs/0903.4035, 2009.

[22] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: mining a social network with negative edges. In *WWW'09*, pages 741–750, 2009.

[23] C. Lampe, E. W. Johnston, and P. Resnick. Follow the reader: filtering comments on slashdot. In *CHI'07*, pages 1253–1262, 2007.

[24] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg. Predicting positive and negative links in online social networks. In *WWW'10*, pages 641–650, 2010.

[25] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg. Signed networks in social media. In *CHI'10*, pages 1361–1370, 2010.

[26] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N. S. Glance. Cost-effective outbreak detection in networks. In *KDD'07*, pages 420–429, 2007.

[27] P. Massa and P. Avesani. Controversial users demand local trust metrics: An experimental study on epinions.com community. In *AAAI'05*, pages 121–126, 2005.

[28] F. McSherry. A uniform approach to accelerated pagerank computation. In *WWW'05*, pages 575–582, 2005.

[29] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, Nov. 1999.

[30] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007.

[31] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD'09*, pages 807–816, 2009.