Constructing Negative Links from Multi-facet of Social Media

Lin Li[1], YunYi Yan[2], LiBin Jia[1], and Jun Ma[3*]
[1] School of Computer Science, Zhengzhou University of Aeronautics
Zhengzhou, 450064 - China
[e-mail: lilin, jialibin@zzia.edu.cn]
[2] School of Aerospace Science and Technology, Xidian University
Xi'an, 710071 - China
[e-mail: yyyan@xidian.edu.cn]
[3] School of Information and Electronics, Beijing Institute of Technology,
Beijing, 100081 - China
[e-mail: junma@bit.edu.cn]
*Corresponding author: Jun Ma

Received August 20, 2016; revised December 28, 2016; accepted January 22, 2017; published May 31, 2017

Abstract

Various types of social media make the people share their personal experience in different ways. In some social networking sites. Some users post their reviews, some users can support these reviews with comments, and some users just rate the reviews as kind of support or not. Unfortunately, there is rare explicit negative comments towards other reviews. This means if there is a link between two users, it must be positive link. Apparently, the negative link is invisible in these social network. Or in other word, the negative links are redundant to positive links. In this work, we first discuss the feature extraction from social media data and propose new method to compute the distance between each pair of comments or reviews on social media. Then we investigate whether we can predict negative links via regression analysis when only positive links are manifested from social media data. In particular, we provide a principled way to mathematically incorporate multi-facet data in a novel framework, *C*onstructing *N*egative *L*inks, CsNL to predict negative links for discovering the hidden information. Additionally, we investigate the ways of solution to general negative link predication problems with CsNL and its extension. Experiments are performed on real-world data and results show that negative links is predictable with multi-facet of social media data by the proposed framework CsNL. Essentially, high prediction accuracy suggests that negative links are redundant to positive links. Further experiments are performed to evaluate coefficients on different kernels. The results show that user generated content dominates the prediction performance of CsNL.

Keywords: Social network, negative link, non-linear regression, link prediction, multi-kernel learning

https://doi.org/10.3837/tiis.2017.05.010 **ISSN : 1976-7277**

1. Introduction

Social network sites make people share their idea and connect with each other in multiple ways [1]. Users post their knowledge on Wikipedia, they share their personal experience on Tripadvisor, Fackbook, Twitter, and they play network game together on Battle.net or other platforms. These social media data demonstrate several characteristics, such as large-scale, diversity and noisy. There is one more important characteristic is that social media data exhibit extremely imbalance on the sentimentality which means users tend to be benevolent to others rather than expressing their dislike or offensive to others. This scenario can be explain as positive links dominate the social connections and there are some missing links in social network. However, those negative links can add significant value over positive links in practical applications [2]. For example, constructing negative links may identify the small unique group as anomaly from large network for commerce or security. A small number of negative links can improve the performance of recommendation systems. And negative links themselves explicit special structure of the social network. Predicting or constructing negative links in social network is an imperative for better designing and using social network. Unfortunately, the negative links are invisible to us on most social media, and to discover them entails extremely challenges.

In signed network, there are two active research topics, positive link prediction and negative link prediction. Positive link prediction is about inferring new positive links with positive network. Usually, network topological features are extracted from existing positive links to predict new ones such as common neighbors, shortest path length, and Pagerank scores [3]. [4] model the users' correlation between their preference to find close friend in network. [5] is using PageRank to calculate the status scores for positive links prediction.

Negative link prediction can be converted to the problem of regression or classification if only positive link available [6][7][8]. Unsupervised methods is proposed for variation of the link prediction [3]. However, some researcher find that without considering negative links, only positive links can be biased [9][10]. In application, using positive links may lead to over-estimation of the effects of erroneous recommendations [11][12].

Thus, predicting negative links using available sources require novel algorithms in supervised or unsupervised mode. Matrix factorization techniques are chosen for building recommendation systems [13] [14] [15]. Leskovec et al. [16] evaluates the generalization performance across social media site with learned link classifiers. Unsupervised learning problems is transformed into supervised learning problem for improving the prediction accuracy [17][18]. Jiliang Tang et al. predicted distrust from only trust and incorporate social network theory, such as social status, balance theory etc. for negative link prediction [10].

Our work is similar to [19] where SVM kernel is used as classification for negative link prediction. The framework CsNL we proposed in this work takes advantage of the multi-facet of the social media data. CsNL incorporates different information from network itself and content users posted into a multi-kernel model. Not only the prediction accuracy is improved, but the contribution from different facet of data can be estimated which prove valuable in understanding the component of social media.

The major contributions of this paper are summarized below:

• Providing a mechanism to measure the positive link and content-centric social media data for negative links construction;

• Proposed a new framework CsNL to predict the negative links from multi-facet information in social media data; and

• Conducting experiments on practical datasets to demonstrate the effectiveness and generalization of the proposed framework and evaluate the contribution of data in dominate feature space.

The rest of the paper is organized as follow. In Section 2, we briefly analyze the data sets and define the measurement for prediction. In Section 3, we formally define the negative links prediction in term of linear regression. In Section 4, we propose a new framework CsNL with a mathematical formulation for constructing negative links from multi-facet of social media data, and we discuss the ways of solution to general negative links prediction problems with CsNL and its extension. Experimental analysis is presented in Section 5. Section 6 reviews other related works. Conclude in Section 7 with future work.

2. Data Sets Analysis

The social media sites such as Epinions and Slashdot [24] [25] are recent availability of signed networks which draw great attention on signed network researhing. In this study, Epinions and Slashdot social networks are used as our dataset for analysis. We collect two dataset from these two sites. Epinions is a product review website where users can post reviews and create a link to others with positive or negative attitude which is an explicit identification of link sign. Slashdot is a news website working in the similar mechanism to Epinions. User can tag each other as “friends” or “foes” which are identification of link sign. CsNL is trying to predict negative links in these networks. All the negative links can be used as ground-truth for evaluation. The statistics of the dataset is summarized in **Table 1**.

Table 1. Statistics of the Epinions and Slashdot

	Epinions	Slashdot
Number of Users	15,230	7,367
Number of Positive Links	255,315	62,347
Number of Negative Links	51,670	18,912
Number of Posts	596,468	280,246
Number of Positive Rates	6,235,673	1,693,234
Number of Negative Rates	153,236	39,498

2.1 Feature Extraction

Social media data is noisy and highly diversity, it is challenging, if not impossible, to accurately predict the negative links without a good feature representation. It is thus desirable to develop a systematical feature representation approach to effectively characterize the nature of negative links.

If the negative links are invisible to users, there must be some relation between positive links and negative links. This means negative links are redundant to positive links. In order to find this relationship, we need to properly evaluate the properties of the positive links and extract most useful features for machine learning approach, here we employ regression analysis. We first follow similar procedure described in [20] [19] to extract two types of

features which are users' features and users' pair features.

- Users' features are based on the degree from perspective of graph. We extract user's indegree and outdegree to denote the interactions. Other features are the number of triads that each user involved, the number of reviews or comments that user posted, the number of user ratings, and the number of positive ratings or negative ratings, for example, in Epinions, ratings larger than 3 as positive, those lower than 3 as negative.
- Users' pair features are extracted from the properties of a pair of users. These features are the number of positive or negative ratings, Jaccard coefficients of indegree or outdegree of user pair, the number of triad involving edge of user pair.

The above two features subsets are mainly based on properties from graph. There is another important content, user's reviews and comments that should be considered to extract useful features for better description of datasets. The contents users created may be short and have lots of abbreviation, so we extract feature using approach in [21]. Then we use radial bias function to compute the distance between two contents as a distance between these two users. Suppose each post c_i is represented as a feature vector f_i, the distance between two users is defined as:

$$D_{u_i,u_j}(f_i, f_j) = \frac{\|f_i - f_j\|}{\|f_i\|\|f_j\|}$$

Unfortunately, if the user did not post, the distance with this user can't be measured in term of the contents. Here we put the interaction between u_i and u_j into consideration when defining the distance. There are two scenarios exist: one user has post and another have no post, both users have no post. Users without posts can have a rating of positive or negative. Thus we have other 4 types of possible distances defined as below:

u_i has post, u_j has no post with positive rating to u_k.

$$D_{u_i,u_j}(f_i, f_k) = \frac{\|f_i - f_k\|}{\|f_i\|\|f_k\|}$$

u_i has post, u_j has no post with negative rating to u_k.

$$D_{u_i,u_j}(f_i, f_k) = 1 - \frac{\|f_i - f_k\|}{\|f_i\|\|f_k\|}$$

u_i with negative rating to u_k, u_j with positive rating to u_p.

$$D_{u_i,u_j}(f_k, f_p) = 1 - \frac{\|f_k - f_p\|}{\|f_k\|\|f_p\|}$$

u_i and u_j with same kind of ratings to their interaction users, u_k and u_p.

$$D_{u_i,u_j}(f_k, f_p) = \frac{\|f_k - f_p\|}{\|f_k\|\|f_p\|}$$

If the user has more than one post or rating, we randomly choose one of post or interaction to calculate the distance. Because it is more likely that the users share the similar opinions to the product or service. The user features, pair features and content features can be description of data from three sources. Each of them can be used for machine learning based approaches. The combination of them gives a comprehensive description which lead to a better regression performance.

3. Problem Statement

Let $\boldsymbol{u}=\{u_1, u_2,...,u_N\}$ be the set of N users, $\mathrm{A}^p \in R^{N\times N}$ represents positive links where $\mathrm{A}_{ij}^p=1$ if u_i has a positive link to u_j and $\mathrm{A}_{ij}^p=0$ otherwise. Let $\boldsymbol{p}=\{p_1, p_2,..., p_M\}$ be the set of M pieces of reviews (comments). We use $B\in R^{N\times M}$ to denote the user-review relationships where $B_{ij}=1$ if p_j is posted by user u_i, and $B_{ij}=0$ otherwise. Reviews express the opinion to the service or products. If user u_j makes comments on that reviews, we can obtain opinion of support or disagree with off-the-shelf opinion mining tools. User u_j's opinion expresses directly from rating if there is no comments made by user. With or without comments, we can form a user-user relations matrix $L_{ij}\in R^{N\times N}$ where $L_{ij}=1$, $L_{ij}=-1$ and $L_{ij}=0$, if user u_i gives or rate positive, negative or neutral (or no) opinions respectively, between u_i and u_j.

Now we can formulate the problem of negative link construction with only positive links. In this formulation, we try to evaluate that if negative links are redundant to positive links. By extracting topological features from positive links, we should find a transformation matrix $\mathbf{W}\in R^{N\times N}$ to transfer positive links to negative links, i.e., $\mathbf{W}\mathbf{A}_p\approx\mathbf{A}_n$. Apparently a global optimal solution of $\mathbf{W}$ can be obtained via the following linear regression problem as:

$$\min_{\mathbf{W}}\left\|\mathbf{W}\mathbf{A}_p-\mathbf{A}_n\right\|_F^2+\alpha\left\|\mathbf{W}\right\|_F^2 \tag{3.1}$$

where the term $\left\|\mathbf{W}\right\|_F^2$ is introduced to avoid overfitting. With $\mathbf{W}$, given all positive links about a user u_i, we can predict u_i's negative links. Essentially, high prediction accuracy can prove our evaluation that negative links are redundant to positive links. As we know, social media data are noisy and diversity. The prediction of negative from only positive links shows poor performance using linear model with each subset features or combination of three. **Table 2** shows the prediction performance with linear regression model and random guessing in term of accuracy. We find that performance of linear model is no better than random guessing.

Table 2. Prediction accuracy using Random method and linear regression method.

Algorithm	Epinions	Slashdot
Random	0.0002	0.0003
Linear regression	0.0003	0.0003

In Section 2, we extract three sets of features from social media data. These features give multi-facet description of the characteristics of social media. From another perspective, each feature subset means data comes from one source data. So social media is from three sources of data. As we already know that, simply combining three features subsets for linear regression gives a prediction performance no better that random guessing (see **Table 2**), we think there can be other way to combining them. In addition, the three features subsets are raising another interesting problem. Which does subset of features dominate the performance of negative link prediction on different data?

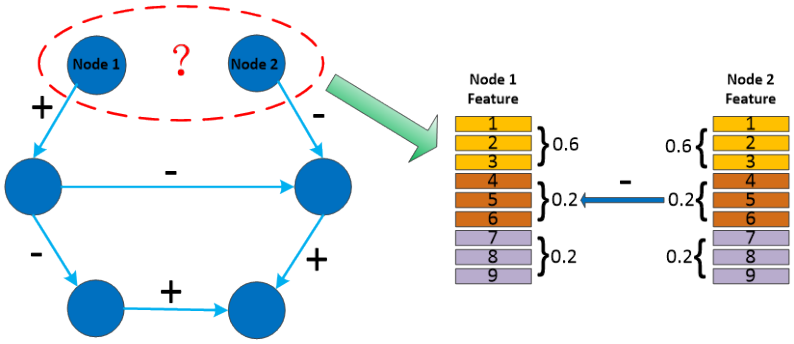


Fig. 1. The proposed framework of NsCL

In **Fig. 1**, we illustrate the differences for the existing variations of the problem as follows:

— Constructing the links and predicting the sign of these links at the same time. The existing variation assumes the links already existed and only predict the sign.

— Identifying the contribution of the feature subset for the prediction of the negative link. The existing variation is without the ability for the estimation of the contribution.

4. The Framework of CsNL

An SVM framework can be used for regression with kernel, it has better generalization capabilities than other learning models. The three features subsets extracted can be considered as three sources of dataset. So our framework of CsNL is based on multiple kernels SVM which taking advantage of the multiple sources of data to perform nonlinear regression analysis. However, the problem we consider in the first place is about only positive links existed. Suppose the users have negative interactions relationship, we briefly analyze that the negative interactions, which are mainly from lower ratings, do prove the existence of negative links. Note that the negative links in datasets are served only as a ground-truth for evaluation. This assumption is proved in [19] that users with negative interactions are likely to have negative links. So next, we will first give details about negative links construction, and then we formulate the problem and discuss the solution.

4.1 Negative Links Construction

The content (reviews) the users posted expresses the opinion of the user u_i according to personal experience of the service or product. If another user u_j supports this opinion, user can make positive comments or give a high rate. If user u_j dislikes service or product, the negative comments can be made or low rate is given. This negative interaction from u_i to u_j manifests the possible existence of negative link from u_j to u_i.

Hypothesis test is designed on Epinions to help us make an inference about the existence of negative links at the desired level of confidence. We follow the similar procedure [19] to conduct a two-sample t-test. We randomly select users' pair with and without negative interactions as sample *a* set and simultaneously, another sample set *b* is collected according to existence of real negative links in sample set *a*. The null hypothesis $H_0: a \leq b$ is strongly accepted (Note: refer to [19] for detail). Thus, it is fair to state that negative interaction between users' pair is a manifested indication of the negative links.

Taking negative interactions as negative links is a straight way to construct the negative

links, which can be used as ground truth for evaluation of the prediction performance from only positive links. The datasets we used are from Epinions and Slashdot which have negative links ready for analysis. The way we provide for negative links construction definitely can be used for those with only positive links available.

4.2 Formulation

From Section 3 we know, negative link prediction can be formulated as regression problem. Eq. (3.1) is a multi-targets linear regression model and do not exploit nonlinearity. We also know that social media data comprised of three set of features, which demonstrate the noisy, and highly diversity of the dataset. So the nonlinearity and multi-sources data should be put into consideration when model the data. In this work, we take advantage of multi-kernel support vector regression model to do prediction. We already define the feature sets and distance between user pair. Apparently, there are three kernels in model. Since kernels are from different feature set, they contain different information, which may improve the prediction performance.

Let $\chi = \{x_1, x_2, ..., x_N\}$ be training examples. Each x_i is a vector comprised of three feature sets $\mathbf{f}^1, \mathbf{f}^2, \mathbf{f}^3$. Technically, we have three sets of examples represented as $x_i^m, m = 1,2,3$ with dimension of size of $\mathbf{f}^i, i = 1,2,3$. The problem we discuss here is two-class learning problem whereby one of the classes, referred to as positive links class, is well-sampled. This is what we know one-class classification [39]. The goal of this kind of classification is to construct a decision surface around the examples from positive links class in order to distinguish between positive links and negative links.

By taking the inspiration of SVM (Support Vector Machine), the positive links samples can be separated by a hyperplane from feature space, $\mathbf{f}^1, \mathbf{f}^2, \mathbf{f}^3$. Now we can introduce multi-SVM to solve negative links predication problem. The primal problem is thus:

$$\begin{aligned} &\min \ \frac{1}{2}\left(\sum_{j=1}^{m} d_j \left\|\mathbf{w}_j\right\|_2\right)^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{w.r.t.} \ \ \mathbf{w} \in R^{f_1} \times R^{f_2} \times R^{f_3},\ \xi \in R_+^n,\ b \in R \\ &\text{s.t.} \ \ y_i\left(\sum_j \mathbf{w}_j^T x_{ji} + \mathrm{b}\right) \geq 1 - \xi_i,\ \forall i \in \{1, \ldots, \mathrm{n}\} \end{aligned} \tag{4.1}$$

Usually, the decision surfaces $\mathbf{w}_j, j = 1,2,3$, cannot be found in original feature space. Thus under kernel SVMformulation, the linear discriminant with the maximum margin in the feature space is induced by the mapping function Φ. The resulting discriminant function is:

$$f(x) = \omega^T \Phi(x) + \beta \tag{4.2}$$

where ω is weight in kernel feature space $\Phi(x)$. β can be soft-margin regulization term or constant. Here β is a constant for simplifying model.

Usually Eq.(4.2) is solved through the Lagrangian dual optimization formulation:

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{\text{Max}} \ \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(x_i, x_j) \\ &\text{s.t.} \ \ \sum_{i=1}^{N}\alpha_i y_i = 0 \\ &\qquad C \geq \alpha_i \geq 0 \quad \forall i \end{aligned} \tag{4.3}$$

where $k(x_i, x_j) = <\Phi(x_i), \Phi(x_j)>$ is the kernel function, and $\boldsymbol{\alpha}$ is the vector of dual variables

corresponding to each separation constraint. Using the KTT condition, the discriminant function can be written as:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + \beta \tag{4.4}$$

In Eq. (4.4), k is one kernel function. Here we employ linear combination of multi-kernels as

$$\mathrm{K}(x_i, x_j) = \sum_{m=1}^{P} \sigma_i k_m(x_i^m, x_j^m) \tag{4.5}$$

where $p = 3$ in our model because of three feature sets from data. Combining three kernels is useful and complementary by kernels themselves because each kernel gives a presentation of one perspective of data and combination of perspective achieve better regression or classification accuracy than each of the kernels [22]. According to [23], we can have the quadratically-constrained quadratic program (QCQP) dual formulation as:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_1 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i a_j y_i y_j \mathrm{K}(x_i, x_j) \\ &\text{s.t.} \sum_{i=1}^{N} \alpha_i y_i = 0 \\ &C \geq \alpha_i \geq 0 \end{aligned} \tag{4.6}$$

The coefficient σ_i is determined by training data from combining kernel as

$$\mathrm{K}(x_i, x_j) = \sum_{m=1}^{P} \sigma_m(x_i \mid D) k_m(x_i^m, x_j^m) \sigma_m(x_j \mid D)$$

where $\sigma_m(x_i \mid D)$ is defined as

$$\sigma_m(x_i \mid D) = \frac{\exp(< \nu_m, \Phi(\mathrm{x}) > + \nu_{m0})}{\sum_{l=1}^{p} \exp(< \nu_l, \Phi(\mathrm{x}) >) + \nu_{l0}}$$

QCQP approach gives significantly better results than any single kernel and the unweighted sum of kernels. We can use gradient-descent method to train model with Function (4.6) which denoted as $J(\mathbf{v})$ [23].

$$\frac{\partial J(\mathbf{v})}{\partial \upsilon_{m0}} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{P} \alpha_i \alpha_j y_i y_j \sigma_k(x_i) \mathrm{K}(x_i, x_j) \sigma_k(x_j) (1 - \sigma_m(\mathrm{x}_i) + 1 - \sigma_m(\mathrm{x}_j)) \tag{4.7}$$

$$\frac{\partial J(\mathbf{v})}{\partial \upsilon_m} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{P} \alpha_i \alpha_j y_i y_j \sigma_k(x_i) \mathrm{K}(x_i, x_j) \sigma_k(x_j) (x_i[1 - \sigma_m(\mathrm{x}_i)] + x_j[1 - \sigma_m(\mathrm{x}_j)]) \tag{4.8}$$

Following the training process in $J(\mathbf{v})$ [23], we update $\mathrm{K}(x_i, x_j)$ each step then update υ_{m0} and υ_m until convergence.

4.3 Discussion

In social media, researchers are mainly working on passive way to study user-generated content and interactions. Finding and understanding the negative links in social media is like a two-step way that we need to first locate negative links, which are invisible to us, and then we extract meaningful information which is value to us. In order to accomplish these two steps, we need to answer three research questions.

Question 1: are negative links redundant to positive links?

As we discussed above, the negative links can be constructed using only positive links by:

$$\min_{\mathbf{W}} \left\| \mathbf{W}\mathbf{A}_p - \mathbf{A}_n \right\|_F^2 + \alpha \left\| \mathbf{W} \right\|_F^2 \tag{4.9}$$

If the negative links are redundant to positive links, some topoplgical features from positive links might be sufficient to indicate the existence of negative links. Essentially, by employing highly-craft prediction method as in our work, those features can be further grouped according to the prediction contribution.

Question 2: how to predict the negative links with co-existence of positive links and negatives?

In some applications, there are some negative links available for predication. We can extend Eq.(3.1) to take advantage of those labeled samples, as

$$\min_{\mathbf{W}} \left\| \mathbf{W}\mathbf{A}_p - \mathbf{A}_n \right\|_F^2 + \alpha \left\| \mathbf{W} \right\|_F^2 + \beta \left\| \mathbf{W}^{index} \mathbf{A}_p^{index} - \mathbf{A}_n^{index} \right\|_F^2 \tag{4.10}$$

where $\mathbf{A}_p^{index}$ and $\mathbf{A}_n^{index}$ are the labeled samples. We can solve this optimization problem using SVM formulation by adding one constraint for each samples in data set. The constraint calculates the prediction error as if the positive links were negative links. Thus we reach to the semi-supervised SVM as:

$$\begin{aligned} &\min \ \frac{1}{2}\left(\sum_{j=1}^{m} d_j \left\| \mathbf{w}_j \right\|_2\right)^2 + C\sum_{i=1}^{n} \xi_i \\ &\text{w.r.t.} \quad \mathbf{w} \in R^{f_1} \times R^{f_2} \times R^{f_3},\ \xi \in R_+^n,\ b \in R \\ &\text{s.t.} \quad y_i\left(\sum_j \mathbf{w}_j^T x_{ji} + \mathrm{b}\right) \geq 1 - \xi_i,\ \forall i \in \{1, \ldots, \mathrm{n}\} \\ &\qquad -\left(\sum_j \mathbf{w}_j^T x_{ji} - \mathrm{b}\right) \geq 1 - \eta_i \end{aligned} \tag{4.11}$$

This is an extension of CsNL in our work, which can be solved using approach described in [23].

Question 3: What is the added value of negative links?

The negative links can be obtained by framework CsNL and its extension under different scenarios. Now we can apply negative links in some real-world applications such as recommendation and classification.

Recommendation: Let $v = \{v_1, v_2, ..., v_I\}$ be items set. R_{ij} is the rating score. We can take positive links $\mathbf{A}_p$ and negative links $\mathbf{A}_n$ as sources for recommendations. For example, positive links can be used to find relevant items and negative links to eliminate some items otherwise to be recommended. Let $\mathrm{O} = \{\langle u_i, v_j \rangle \mid \mathrm{R}_{ij} \neq 0\}$ be the set of observed ratings and $M = \{\langle u_i, v_k \rangle \mid \mathrm{R}_{ik} \neq 0\}$ be the set of missing ratings. Then we reach to the recommendation problem as:

$$f : \{O, A_p, A_n\} \rightarrow \{\hat{R}_{ik}, \text{for} \langle u_i, v_k \rangle \in M\} \tag{4.12}$$

Appearently, the problem described in Eq. (4.12) is the extension of CsNL described in Eq. (4.11). Other works[15][40] on recommendation are only putting positive links into consideration.

Classification: We can build up the user-post links with distance we defined in Section 2. Let U be this distance matrix and F be features set. X denotes feature-value of each data sample. Y is the labe indicatior matrix. $\mathbf{A}_p$ and $\mathbf{A}_n$ be the known positive links and negative links. The classification problem is to predict $\hat{Y}$ for labeled data $\hat{X}$. Since the framework CsNL we proposed in this work is working on one-class sample. We can model the problem into one-class classification problem as Eq. (4.2) described. Then the samples not in this class can be treated as another class. The classification problem requires further study according to the application, such as node classification [41].

5. Experimental Analysis

In this section, we first perform the experiments for quantity analysis of the proposed framework of CsNL. Then we report the coefficient of kernels' combination which represents the contribution of various features to the performance.

Because we predict negative links from only positive links, this formulation can be viewed as one-class classification. So we employ three metrics for evaluation, they are precision, recall and F1-measure defined as:

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Several kernel functions are used for construction of CsNL respectively. They are Gaussian kernel $k(x_i, x) = \exp(-\frac{\|x_i - x\|^2}{2\sigma^2})$, cosine kernel $k(x_i, x) = \frac{\pi}{4}\cos(\frac{\pi}{2}\|x_i - x\|^2)$, logistic kernel $k(x_i, x) = \frac{1}{e^{\|x_i - x\|^2} + 2 + e^{-\|x_i - x\|^2}}$ and sigmoid kernel $k(x_i, x) = \frac{2}{\pi}\frac{1}{e^{\|x_i - x\|^2} + e^{-\|x_i - x\|^2}}$.

Since constructing negative links can be viewed as problem of one-class classification, we need to choose binary classification function for discriminant analysis. Because logistic sigmoid function can introduce non-linearity to the model and, from the real-world data analysis, we know that the distance between pair of samples is within certain range. The logistic sigmoid function is working well as resulting discriminant function in Eq. (4.1) for CsNL.

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{5.1}$$

The first experiment is to access the performance of CsNL on different kernel functions. By using different kernel function, CsNL model is formulated according to problem (4.5). After the training according to Eq. (4.6), Eq. (4.7) with two datasets respectively, then we perform the prediction. **Table 3** reports the experiments results on two datasets.

Table 3. Prediction performance using different kernel function on two datasets.

Kernel function	Epinions			Slashdot		
	Precision	Recall	F1	Precision	Recall	F1
Gaussion	0.2956	0.3201	0.3073	0.2203	0.2245	0.2224
Cosine	0.2901	0.3192	0.3040	0.2191	0.2135	0.2162
Logistic	0.2842	0.3110	0.2970	0.2031	0.2105	0.2067
Sigmoid	0.2801	0.3119	0.2951	0.2010	0.2127	0.2067

The first row of the **Table 3** shows the best performance of NsCL when using Gaussion kernel. The features we defined in our study are all real number. Gaussion kernel makes a good fitting to our model.

In the next experiments, we compare the performance of NsCL with other based-line methods. The baseline methods are as follow:

- **Random:** The labels of the links are randomly guessed.
- **Single SVM kernel methods:** Nonlinear prediction model which combining three feature sets into one for constructing one kernel under SVM setting.
- **NeLP** [19]: multi-kernel with fixed coefficient.

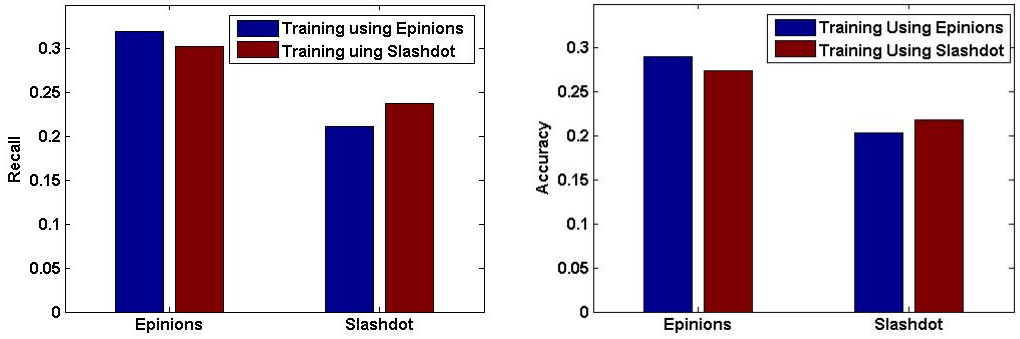
Discriminant function in Eq. (5.1) is used for all the base-line methods. Gaussion kernel is used for construction of each methods. We set the parameter for NeLP as reported in [19]. The experiments results are reported in **Table 4**.

Table 4. Prediction performance using different prediction methods on two datasets.

Methods	Epinions			Slashdot		
	Precision	Recall	F1	Precision	Recall	F1
Random	0.0002	0.0005	0.0003	0.0004	0.0005	0.0004
Single SVM	0.2539	0.3768	0.3034	0.1992	0.1936	0.1964
NeLP	0.2915	0.3810	0.3303	0.2131	0.2105	0.2118
NsCL	0.2956	0.3901	0.3363	0.2203	0.2245	0.2224

The first row in **Table 4** is the results of random guessing. Without doubt about it, the performance is extremely low. The second row of single SVM is actually a special case of NsCL with one kernel employed. This method takes the feature set as whole for the kernel building. The third row of NeLP is also a special case of NsCL with pre-fixed coefficients on each kernel. The drawback of this method is that we have to set these coefficient in advance. From the results in fourth row, the precision and recall are all slight lower than NsCL. The premier characteristic of NsCL is that the coefficient is evaluated from the datasets themselves. This makes the NsCL exhibit high robustness to the diversity datasets.

As stated in [19], the prediction model trained from social media data can be consider as a generalization model for broadly available social media data. So our next experiments are followed the similarly setting in [19] to evaluate the performance of CsNL. We first train CsNL with Epinions and perform classification on Epinions and Slashdot. Then we train CsNL with Slashdot and perform classification on Slashdot and Epinions.



(a) Recall (b) Accuracy

Fig. 2. The prediction performance using different dataset for training and testing.

As **Fig. 2** shows, CsNL remain relative high classification performance when testing on different dataset. These result prove the good generalization when CsNL are working as a classifier with different training data.

Since framework of CsNL is multi-kernel model, the coefficient on each kernel can be the evaluation of kernel contribution to prediction. The kernel coefficient is computed from

training data. The last experiment is to evaluate the dominance of the source of data in term of the kernel coefficient. We report the results on two datasets in **Fig. 3**.

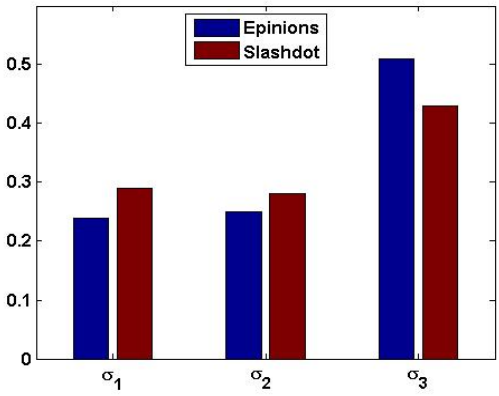


Fig. 3. Coefficients on different kernel

Apparently, kernel from distance defined in subsection 2.1 achieves high coefficient. This result means the contents user posted dominate the performance of the prediction. This finding reveals that the features we define in subsection 2.1 provide powerful discriminative capability for negative prediction. And these features bridge the gap for the sentimentality discovery even the users have no comments or reviews.

6. Other Related Work

In this section, we briefly review works which is related to social network mining and its applications.

The power-law distributions is a well known propertity of the social network [26]. J. Tang et.al. reported that in signed network positive links are denser thatn negative links [27]. Clustering coefficient is another property introduced by [28]. This property cannot be applied to signed network mining, especially for negative links. There are two properties, reciprocity and transitivity, which are widly used in positve link predictions [29][27]. Balance theory [30] and power-law distribution are very useful properties for many social mining applications, such as community discovery [31], information diffusion [32] and data classification [33] or clustering [34].

The techniques in community discovery can be directly applied to recommendation system [35]. Information diffusion is applied to marketing [32][36]. Classification and clustering are techniques that are more general. Research efforts have mainly focus on signed network, especially on negative link networks [37][38].

7. Conclusion

Negative links provide valuable information to social media. However, most social media do not support user to specify negative links. It is necessary to use positive links and other information such as review and rating to inference the existence of the negative links. Social media data is diversity and in large scale, this make the problem of negative link prediction challenging. In this paper, a new framework of CsNL is proposed which takes multi-facet of

social media data in to consideration to construct negative links. We first analyze the feature of dataset and define a new measurement for nonlinear analysis. Then CsNL based on multi-kernel SVM is built for exploit the information from positive links for negative links prediction. Experiments are performed on real-world datasets. Experimental results prove the effective and generalization of CsNL. Moreover, the framework evaluation demonstrates the contribution for prediction from each source of data.

Future work will be on the investigation of incorporation model with light negative information existed which can further improve the prediction accuracy. We plan to put these models into practical application such as recommendation systems or social network security system.

References

[1] Andreas M Kaplan and Michael Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business horizons*, 53(1):59-68, 2010. Article (CrossRef Link)

[2] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu, "Social Media Mining: An Introduction," *Cambridge University Press*, 2014. Article (CrossRef Link)

[3] David Liben-Nowell and Jon Kleinberg, "The link prediction problem for social networks," in *Proc. of 12th International Conference on Information and Knowledge Management*. Article (CrossRef Link)

[4] Aditya Krishna Menon and Charles Elkan, "Link prediction via matrix factorization, *Machine Learning and Knowledge Discovery in Databases*, pp. 437-452, Springer, 2011. Article (CrossRef Link)

[5] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, The pagerank citation ranking: Bringing order to the web, 1999. Article (CrossRef Link)

[6] Xiaoli Li and Bing Liu, "Learning to classify texts using positive and unlabeled data," *IJCAI*, volume 3, pp. 587-592, 2003. Article (CrossRef Link)

[7] Charles Elkan and Keith Noto, "Learning classi_ers from only positive and unlabeled data," in *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213-220. ACM, 2008. Article (CrossRef Link)

[8] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla, "New perspectives and methods in link prediction," in *Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 243-252. ACM, 2010. Article (CrossRef Link)

[9] Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins, "Propagation of trust and distrust," in *Proc. of the 13th international conference on World Wide Web*, pp. 403-412. ACM, 2004. Article (CrossRef Link)

[10] Jiliang Tang, Xia Hu, and Huan Liu, "Is distrust the negation of trust? the value of distrust in social media," in *Proc. of ACM Hypertext conference*, 2014. Article (CrossRef Link)

[11] Yanhua Li,Wei Chen, YajunWang, and Zhi-Li Zhang, "Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships," in *Proc. of the sixth ACM international conference on Web search and data mining*, pp. 657-666. ACM, 2013. Article (CrossRef Link)

[12] Vincent Traag, Yurii Nesterov, and Paul Van Dooren, "Exponential ranking: Taking into account negative links," *Social Informatics*, pp. 192-202, 2010. Article (CrossRef Link)

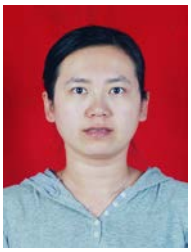
[13] Hao Ma, Haixuan Yang, Michael Lyu, and Irwin King, "Sorec: social recommendation using probabilistic matrix factorization," in *Proc. of the 17th ACM conference on Information and knowledge management*, pp. 931-940. ACM, 2008. Article (CrossRef Link)

[14] Yehuda Koren, Robert Bell, and Chris Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, 42(8):30-37, 2009. Article (CrossRef Link)

[15] Hao Ma, Dengyong Zhou, Chao Liu, Michael R Lyu, and Irwin King, "Recommender systems with social regularization," in *Proc. of the fourth ACM international conference on Web search and data mining*, pp. 287-296. ACM, 2011. Article (CrossRef Link)

[16] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. of the 19th international conference on World wide web*, pp. 641-650. ACM, 2010. Article (CrossRef Link)
[17] Deng Cai, Chiyuan Zhang, and Xiaofei He, "Unsupervised feature selection for multi-cluster data," in *Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 333-342. ACM, 2010. Article (CrossRef Link)
[18] Jiliang Tang and Huan Liu, "Unsupervised feature selection for linked social media data," in *Proc. of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 904-912. ACM, 2012.Article (CrossRef Link)
[19] Jiliang Tang, Shiyu Chang, Charu Aggarwaly and Huan Liu, "Negative Link Prediction in Social Media," *WSDM'15*, February 2–6, Shanghai, China, 2015. Article (CrossRef Link)
[20] J. Leskovec, D. Huttenlocher and J. Kleiberg, "Predicting Positive and Negative Links in Online Social Networks," in *Proc. of WWW 2010*, April 26-30, Raleigh, North Carolina, USA, 2010. Article (CrossRef Link)
[21] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in *Proc. of KDD'04*, August 22-25, Seattle, Washington, USA, 2004. Article (CrossRef Link)
[22] Thorsten Joachims, Nello Cristianini, and John Shawe-Taylor, "Composite kernels for hypertext categorisation," in *Proc. of the 18th International Conference on Machine Learning*, 2001. Article (CrossRef Link)
[23] Mehmet Gonen, Ethem Alpaydin, "Localized Multiple Kernel Learning," in *Proc. of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Article (CrossRef Link)
[24] http://www.epinions.com/
[25] http://slashdot.org/
[26] LA Adamic, BA Huberman, AL Barabási, R Albert, H Heong and G Bianconi, "Power-Law Distribution of the World Wide Web," *Science*, 287(5461):2115, 2000. Article (CrossRef Link)
[27] Jiliang Tang, Xia Hu, and Huan Liu, "2014a. Is Distrust the Negation of Trust? The Value of Distrust in Social Media," in *Proc. of ACM Hypertext conference*. Article (CrossRef Link)
[28] James S Coleman, "Social capital in the creation of human capital," *American journal of sociology*, S95–S120, 1988. Article (CrossRef Link)
[29] Michael Szell, Renaud Lambiotte, and Stefan Thurner, "Multirelational organization of large-scale social networks in an online world," in *Proc. of the National Academy of Sciences 107*, 31, 13636–13641, 2010. Article (CrossRef Link)
[30] Dorwin Cartwright and Frank Harary, "Structural balance: a generalization of Heider's theory," *Psychological review 63*, 5 277, 1956. Article (CrossRef Link)
[31] Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos, "Community detection in social media," *Data Mining and Knowledge Discovery* 24, 515–554, 2012. Article (CrossRef Link)
[32] David Kempe, Jon Kleinberg, and ´Eva Tardos, "Maximizing the spread of influence through a social network," in *Proc. of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 137–146, 2003. Article (CrossRef Link)
[33] Xiaoguang Qi and Brian D Davison, "Web page classification: Features and algorithms," *ACM Computing Surveys (CSUR) 41*, 2, 12, 2009. Article (CrossRef Link)
[34] Mahdokht Masaeli, Jennifer G Dy, and Glenn M Fung, "From transformation-based dimensionality reduction to feature selection," in *Proc. of the 27th International Conference on Machine Learning (ICML-10)*, 751–758, 2010. Article (CrossRef Link)
[35] Rana Forsati, Mehrdad Mahdavi,Mehrnoush Shamsfard, and Mohamed Sarwat, "Matrix Factorization with Explicit Trust and Distrust Side Information for Improved Social Recommendation," *ACM Transactions on Information Systems (TOIS) 32*, 4, 17, 2014. Article (CrossRef Link)
[36] Wei Chen, Yajun Wang, and Siyu Yang, "Efficient influence maximization in social networks," in *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 199–208, 2009. Article (CrossRef Link)

[37] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *Proc. of the 22nd international conference on Machine learning. ACM*, 824–831, 2005. Article (CrossRef Link)
[38] KiriWagstaff, Claire Cardie, Seth Rogers, Stefan Schr¨odl, and others, "Constrained k-means clustering with background knowledge," *ICML*, Vol. 1. 577–584, 2001. Article (CrossRef Link)
[39] Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification 2nd Edition. 2004. John Wiley. Article (CrossRef Link)
[40] Mohsen Jamali, and Martin Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proc. of* the fourth ACM conference on Recommender systems, ACM, 135-142, 2010. Article (CrossRef Link)
[41] Lise Getoor, and Christopher P Diehl, "Link mining: a survey," *ACM SIGKDD Explorations Newsletter*, 7(2):3-12, 2005. Article (CrossRef Link)



Lin Li is a lecturer in School of Computer Science, Zhengzhou University of Aeronautics. Her research interest is on social network mining.

Yunyi Yan is a associate professor in School of Aerospace Science and Technology, Xi'an, China. He received his Ph.D degree in control science and engineering from Xidian University, Xi'an, China, 2008. His research interest is on image processing and embedding system.

Libin Jia is a lecturer in School of Computer Science, Zhengzhou University of Aeronautics. His research interest is on social network security.

Jun Ma is a lecturer in School of Information and Electronics, Beijing Institute of Technology, Beijing, China. He received his Ph.D. degree in control science and engineering from Northwest Polytechnical University, Xi'an, China, 2011. His research interest is on feature selection and social mining.