

Feature Selection with Linked Data in Social Media

Jiliang Tang*

Huan Liu*

Abstract

Feature selection is widely used in preparing high-dimensional data for effective data mining. Increasingly popular social media data presents new challenges to feature selection. Social media data consists of (1) traditional high-dimensional, attribute-value data such as posts, tweets, comments, and images, and (2) linked data that describes the relationships between social media users as well as who post the posts, etc. The nature of social media also determines that its data is massive, noisy, and incomplete, which exacerbates the already challenging problem of feature selection. In this paper, we illustrate the differences between attribute-value data and social media data, investigate if linked data can be exploited in a new feature selection framework by taking advantage of social science theories, extensively evaluate the effects of user-user and user-post relationships manifested in linked data on feature selection, and discuss some research issues for future work.

1 Introduction

The myriads of social media services are emerging in recent years that allow people to communicate and express themselves conveniently and easily, e.g., Facebook¹ and Twitter². The pervasive use of social media generates massive data in an unprecedented rate. For example, users on Twitter are sending 200 million Tweets per day, which is about 200 percent growth in a year³; more than 3,000 photos are uploaded to Flickr⁴ per minutes and more than 153 million blogs are posted per year⁵. The massive, high-dimensional social media data poses new challenges to data mining tasks such as classification and clustering. One conventional approach to handling large-scale, high-dimensional data is feature selection [13].

Feature selection aims to select relevant features

from the high dimensional data for a compact and accurate data representation. It can alleviate the curse of dimensionality, speed up the learning process, and improve the generalization capability of a learning model [14]. The vast majority of existing feature selection algorithms work with “flat” data containing uniform entities (or attribute-value data points) that are typically assumed to be independent and identically distributed (*i.i.d.*). However, social media data differs as its data points or instances are inherently connected to each other. Without loss of generality, Figure 1 presents a simple example of social media data with two data representations. Figure 1(a) has four users (u_1, \dots, u_4) and each user follows some other users (e.g., u_1 follows u_2 and u_4) and has some posts (e.g., user u_1 has two posts p_1 and p_2). We use posts in a loose way to cover posts, tweets, or images. Figure 1(b) is a conventional representation of attribute-value data: rows are posts and columns are features or terms for text posts. Its similarity with social media data stops here. In the context of social media, there is additional information in the form of linked data such as who posts the posts and who follows whom as shown in Figure 1(c). After delineating the differences between attribute-value data and social media data, we now discuss the problem of feature selection with linked data.

In this paper, we investigate issues of feature selection for social media data as illustrated in Figure 1(c). Specifically, we perform feature selection on posts (e.g., tweets, blogs, or images) in the context of social media with link information between user and user or between user and posts. Since conventional feature selection methods cannot take advantage of the additional information in linked data, we proceed to study two fundamental problems: (1) *relation extraction* - what are distinctive relations that can be extracted from linked data, and (2) *mathematical representation* - how to represent these relations and integrate them in a state-of-the-art feature selection formulation. Providing answers to the two problems, we propose a novel feature selection framework (LinkedFS) for social media data. The main contributions of this paper are summarized next.

- Identify the need for feature selection in social media and propose to exploit social correlation

*Computer Science and Engineering Department, Arizona State University, Tempe, AZ. {jiliang.tang,huan.liu}@asu.edu

¹<http://www.facebook.com>

²<http://www.twitter.com>

³<http://techcrunch.com/2011/06/30/twitter-3200-million-tweets/>

⁴<http://www.flickr.com/>

⁵<http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/>

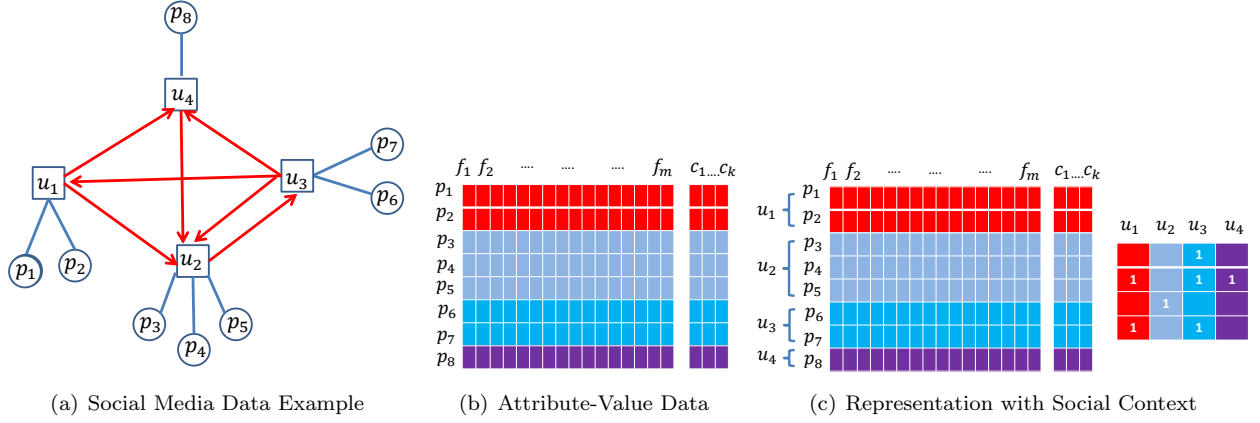


Figure 1: Typical Social Media Data and Its Two Representations

theories and linked data in formulating the new problem of feature selection for social media data;

- Show that various relations can be extracted from linked data guided by social correlation theories and provide a way to capture link information;
- Propose a framework for social media feature selection (LinkedFS) that integrates conventional feature selection with extracted relations; and
- Evaluate LinkedFS systematically using real-world social media data to verify if different types of relations improve the performance of feature selection.

The rest of this paper is organized as follows. The problem of feature selection with linked data in social media is formally defined in Section 2. A new feature selection framework, LinkedFS, is introduced in Section 3 based on social correlation theories. Empirical evaluation is presented in Section 4 with discussions. The related work is reviewed in Section 5. The conclusion and future work are presented in Section 6.

2 Problem Statement

We first give the notations to be used in this paper. Scalars are denoted by low-case letters ($a, b, \dots; \alpha, \beta, \dots$), vectors are written as low-case bold letters ($\mathbf{a}, \mathbf{b}, \dots$), and matrices correspond to bold-face upper-case letters ($\mathbf{A}, \mathbf{B}, \dots$). $\mathbf{A}(i, j)$ is the entry at the i^{th} row and j^{th} column of the matrix \mathbf{A} , $\mathbf{A}(i, :)$ is the i^{th} row of \mathbf{A} and $\mathbf{A}(:, j)$ is the j^{th} column of \mathbf{A} .

Let $\mathbf{p} = \{p_1, p_2, \dots, p_N\}$ be the post set (e.g., p_1 to p_8) where N is the number of posts. Let $\mathbf{f} = \{f_1, f_2, \dots, f_m\}$ denote the feature set where m is the number of features. For each post p_i , $\mathbf{f}_i \in \mathbb{R}^m$ are the set

of feature values where $\mathbf{f}_i(j)$ is the frequency of f_j used by p_i . $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{N_1}\} \in \mathbb{R}^{m \times N}$ denotes the whole dataset \mathbf{p} . We assume that the subset $\{p_1, p_2, \dots, p_{N_l}\}$ is the labeled data where N_l is the number of labeled posts. Then $\mathbf{X} = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{N_l}\} \in \mathbb{R}^{m \times N_l}$ is the matrix for labeled data. Let $\mathbf{c} = \{c_1, c_2, \dots, c_k\}$ denote the class label set where k is the number of classes. $\mathbf{Y} \in \mathbb{R}^{N_l \times k}$ is the class label matrix for labeled data where $\mathbf{Y}(i, j) = 1$ if p_i is labeled as c_j , otherwise zero.

Let $\mathbf{u} = \{u_1, u_2, \dots, u_n\}$ be the user set (e.g., u_1 to u_4) where n is the number of users. \mathbf{F}_i denotes the set of posts from user u_i (e.g., $\mathbf{F}_1 = \{p_1, p_2\}$). We also model the user-user following relationships as a graph with adjacency matrix \mathbf{S} , where $\mathbf{S}(i, j) = 1$ if there is a following relationship from u_j to u_i and zero otherwise (e.g., $\mathbf{S}(:, 1) = [0, 1, 0, 1]^T$). Let $\mathbf{P} \in \mathbb{R}^{n \times N}$ denote user-post relationships where $\mathbf{P}(i, j) = 1$ if p_j is posted by u_i , zero otherwise (e.g., $\mathbf{P}(1, :) = [1, 1, 0, 0, 0, 0, 0, 0]$).

Traditional supervised feature selection aims to select a subset features from m features based on $\{\mathbf{X}, \mathbf{Y}\}$. Different from traditional feature selection problems, our problem with linked data is stated as:

Given labeled data \mathbf{X} and its label indicator matrix \mathbf{Y} , the whole dataset \mathbf{F} , its social context (or social correlations) including user-user following relationships \mathbf{S} and user-post relationships \mathbf{P} , we aim to select K most relevant features from m features on the dataset \mathbf{F} with its social context \mathbf{S} and \mathbf{P} .

3 A New Framework - LinkedFS

Recall the two fundamental problems for feature selection on social media data: (1) relation extraction, and (2) mathematical representation. Their associated challenges are: (a) *what are different types of relations among data instances and how to capture them*, and (b)

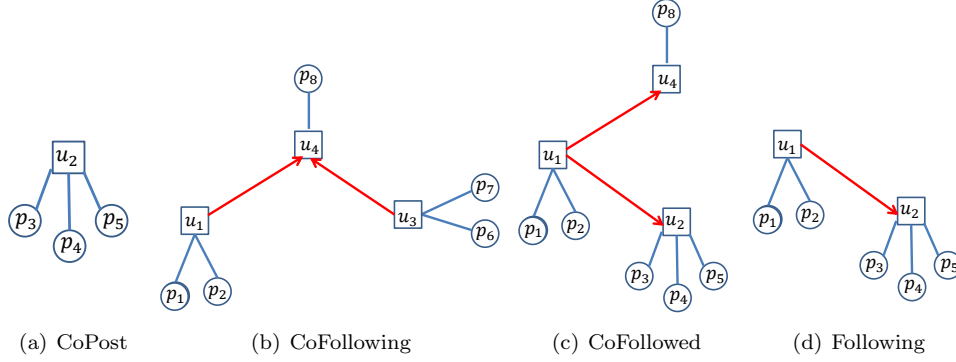


Figure 2: Different Types of Relations Extracted from Social Correlations among Social Media Data

how to model these relations for feature selection. In this section, we discuss how to capture relations from linked data guided by social correlation theories, propose a framework (LinkedFS) of social media data that naturally integrates different relations into a state-of-the-art formulation of feature selection, and turn the integrated formulations to an optimization problem with convergence analysis when developing its corresponding feature selection algorithm.

3.1 Extracting Various Relations Examining Figure 1(a), we can find four basic types of relations from the linked data as shown in Figure 2: (a) a user (u_2) can have multiple posts (p_3, p_4 , and p_5), (b) two users (u_1 and u_3) follow a third user (u_4), (c) two users (u_2 and u_4) are followed by a third user (u_1), and (d) a user (u_1) follows another user (u_2). Social correlation theories such as homophily [17] and social influence [16] can be helpful to explain what these relations suggest. Homophily indicates that people with similar interests are more likely to be linked, and social influence reveals that people that are linked are more likely to have similar interests. Based on these theories that define social correlations among data, we turn the four types of relations into four corresponding hypotheses that can affect feature selection with linked data.

CoPost Hypothesis: This hypothesis assumes that posts by the same user (e.g., $\{p_3, p_4, p_5\}$, in Figure 2(a)) are of similar topics. In other words, the posts of a user are more similar, in terms of topics (say, “sports”, “music”), than those randomly selected posts.

CoFollowing Hypothesis: This hypothesis suggests that if two users follow the same user (e.g., u_1 and u_3 follow u_4 as in Figure 2(b)), their posts, $\{p_1, p_2\}$ and $\{p_6, p_7\}$, are likely of similar topics. Its counterpart in citation analysis is bibliographic coupling [18]: if two papers cite a paper, they are more similar than other

papers that do not share references.

CoFollowed Hypothesis: It says that if two users are followed by the same user, their posts are similar in topics. For example, in Figure 2(c), both users u_2 and u_4 are followed by user u_1 and then their posts $\{p_3, p_4, p_5\}$ and $\{p_8\}$ are of more similar topics. It is similar to the co-citation relation [18] in citation analysis: if two papers are cited by the same paper, they are more similar than other paper that are not.

Following Hypothesis: The hypothesis assumes that one user follows another (e.g., u_1 follows u_2 in Figure 2(d)) because u_1 shares u_2 ’s interests. Thus, their posts (e.g., $\{p_1, p_2\}$ and $\{p_3, p_4, p_5\}$) are more likely similar in terms of topics.

Next, we elaborate how the above four hypotheses can be modeled into a feature selection formulation in our effort to create a new framework for feature selection with linked data.

3.2 Modeling Hypotheses We first introduce a representative feature selection method for attribute-value data based on $\ell_{2,1}$ -norm regularization [15], which selects features across data points with joint sparsity [5].

$$(3.1) \quad \min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \alpha \|\mathbf{W}\|_{2,1}$$

where $\mathbf{W} \in \mathbb{R}^{m \times k}$ and $\|\mathbf{W}\|_{2,1}$ is the $\ell_{2,1}$ -norm of \mathbf{W} , which is defined as follows:

$$(3.2) \quad \|\mathbf{W}\|_{2,1} = \sum_{i=1}^m \sqrt{\sum_{j=1}^k \mathbf{W}^2(i, j)} = \sum_{i=1}^m \|\mathbf{W}(i, :)\|_2$$

This formulation (Eq. (3.1)) is a supervised feature selection method where data instances are assumed to be independent. We now discuss how one can jointly incorporate $\ell_{2,1}$ minimization and different types relations in feature selection with linked data.

CoPost Relation: To integrate this hypothesis into Eq. (3.1), we propose to add a regularization term that enforces the hypothesis that the class labels (i.e., topics in this paper) of posts by the same user are similar. Thus, feature selection with CoPost hypothesis can be formulated as the following optimization problem.

$$(3.3) \quad \min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \alpha \|\mathbf{W}\|_{2,1} + \beta \sum_{u \in \mathbf{u}} \sum_{\mathbf{f}_i, \mathbf{f}_j \in \mathbf{F}_u} \|T(\mathbf{f}_i) - T(\mathbf{f}_j)\|_2^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The parameter α controls the sparseness of \mathbf{W} in rows and β adjusts the contribution from the CoPost relation. Let \mathbf{A} be the copost matrix, which is defined as $\mathbf{A}(i, j) = 1$ if post p_i and post p_j are posted by the same user, and $\mathbf{A}(i, j) = 0$ otherwise. \mathbf{A} can be obtained from the user-post relationship matrix \mathbf{P} , i.e., $\mathbf{A} = \mathbf{P}^\top \mathbf{P}$. Let $T(\mathbf{f}_i) : \mathbb{R}^m \rightarrow \mathbb{R}^k$ be the function to predict the labels of the post p_i , i.e., $T(\mathbf{f}_i) = \mathbf{W}^\top \mathbf{f}_i$. $\mathbf{L}_\mathbf{A} = \mathbf{D}_\mathbf{A} - \mathbf{A}$ is the Laplacian matrix, and $\mathbf{D}_\mathbf{A}$ is a diagonal matrix with $\mathbf{D}_\mathbf{A}(i, i) = \sum_j \mathbf{A}(j, i)$.

THEOREM 3.1. *The formulation in Eq (3.3) is equivalent to the following optimization problem:*

$$(3.4) \quad \min_{\mathbf{W}} \text{tr}(\mathbf{W}^\top \mathbf{B} \mathbf{W} - 2\mathbf{E} \mathbf{W}) + \alpha \|\mathbf{W}\|_{2,1}$$

where $\text{tr}(\cdot)$ is the trace of a matrix. \mathbf{B} and \mathbf{E} are defined as follows:

$$(3.5) \quad \begin{aligned} \mathbf{B} &= \mathbf{X} \mathbf{X}^\top + \beta \mathbf{F} \mathbf{L}_\mathbf{A} \mathbf{F}^\top \\ \mathbf{E} &= \mathbf{Y}^\top \mathbf{X}^\top \end{aligned}$$

Proof. It is easy to verify that the first part of Eq (3.3) can be written as:

$$(3.6) \quad \begin{aligned} &\|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 = \\ &\text{tr}(\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W} - 2\mathbf{Y}^\top \mathbf{X}^\top \mathbf{W} + \mathbf{Y}^\top \mathbf{Y}) \end{aligned}$$

With the definition of \mathbf{F} and $\mathbf{L}_\mathbf{A}$, the last regularization constraint of Eq (3.3) can be written as:

$$(3.7) \quad \begin{aligned} &\sum_{u \in \mathbf{u}} \sum_{\mathbf{f}_i, \mathbf{f}_j \in \mathbf{F}_u} \|T(\mathbf{f}_i) - T(\mathbf{f}_j)\|_2^2 \\ &= \sum_i \sum_j \mathbf{A}(i, j) \|\mathbf{W}^\top \mathbf{f}_i - \mathbf{W}^\top \mathbf{f}_j\|_2^2 \\ &= \text{tr}(\mathbf{W}^\top \mathbf{F} \mathbf{L}_\mathbf{A} \mathbf{F}^\top \mathbf{W}) \end{aligned}$$

Since $\mathbf{Y}^\top \mathbf{Y}$ is constant, the object function can be converted into:

$$(3.8) \quad \begin{aligned} &\text{tr}(\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W} - 2\mathbf{Y}^\top \mathbf{X}^\top \mathbf{W} + \mathbf{Y}^\top \mathbf{Y}) \\ &\quad + \beta \text{tr}(\mathbf{W}^\top \mathbf{F} \mathbf{L}_\mathbf{A} \mathbf{F}^\top \mathbf{W}) \\ &= \text{tr}(\mathbf{W}^\top (\mathbf{X} \mathbf{X}^\top + \beta \mathbf{F} \mathbf{L}_\mathbf{A} \mathbf{F}^\top) \mathbf{W} - 2\mathbf{Y}^\top \mathbf{X}^\top \mathbf{W}) \\ &= \text{tr}(\mathbf{W}^\top \mathbf{B} \mathbf{W} - 2\mathbf{E} \mathbf{W}) \end{aligned}$$

which completes the proof.

CoFollowing Relation: To model this hypothesis, we have to formally define users' topic interests. With the definition of $T(\mathbf{f}_i)$, the topic interests for u_k , $\hat{T}(u_k)$, is defined as follows:

$$(3.9) \quad \hat{T}(u_k) = \frac{\sum_{\mathbf{f}_i \in \mathbf{F}_k} T(\mathbf{f}_i)}{|\mathbf{F}_k|} = \frac{\sum_{\mathbf{f}_i \in \mathbf{F}_k} \mathbf{W}^\top \mathbf{f}_i}{|\mathbf{F}_k|}$$

For this hypothesis, we add a regularization term into Eq. (3.1), reflecting the constraint that two co-following users have similar interested topics. The feature selection formulation with CoFollowing hypothesis is below:

$$(3.10) \quad \begin{aligned} &\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2 + \alpha \|\mathbf{W}\|_{2,1} \\ &\quad + \beta \sum_{u_k \in \mathbf{u}} \sum_{u_i, u_j \in \mathbf{F}_k} \|\hat{T}(u_i) - \hat{T}(u_j)\|_2^2 \end{aligned}$$

where \mathbf{F}_k is the people who follow u_k . Let $\mathbf{F}\mathbf{I}$ be the co-following matrix where $\mathbf{F}\mathbf{I}(i, j) = 1$ if u_i and u_j are following at least one other person (e.g., u_k). $\mathbf{F}\mathbf{I}$ can be obtained from the adjacency matrix \mathbf{S} , i.e., $\mathbf{F}\mathbf{I} = \text{sign}(\mathbf{S}^\top \mathbf{S})$ where the function $\text{sign}(x) = 1$ if $x > 0$ and 0 otherwise.

Let $\mathbf{H} \in \mathbb{R}^{N \times n}$ be an indicator matrix where $\mathbf{H}(i, j) = \frac{1}{|\mathbf{F}_j|}$ if u_j is the author of p_i . Let $\mathbf{L}_{\mathbf{F}\mathbf{I}}$ is the Laplacian matrix defined on $\mathbf{F}\mathbf{I}$.

THEOREM 3.2. *The formulation in Eq (3.10) is equivalent to the following optimization problem:*

$$(3.11) \quad \min_{\mathbf{W}} \text{tr}(\mathbf{W}^\top \mathbf{B} \mathbf{W} - 2\mathbf{E} \mathbf{W}) + \alpha \|\mathbf{W}\|_{2,1}$$

where \mathbf{B} and \mathbf{E} are defined as follows:

$$(3.12) \quad \begin{aligned} \mathbf{B} &= \mathbf{X} \mathbf{X}^\top + \beta \mathbf{F} \mathbf{H} \mathbf{L}_{\mathbf{F}\mathbf{I}} \mathbf{H}^\top \mathbf{F}^\top \\ \mathbf{E} &= \mathbf{Y}^\top \mathbf{X}^\top \end{aligned}$$

Proof. We can see that the first part of Eq (3.3) is the same as that of Eq (3.10). For this part, the proof process is similar as that of CoPost hypothesis. The last

regularization constraint of Eq (3.10) can be written as:

$$\begin{aligned}
(3.13) \quad & \sum_{u_k \in \mathbf{u}} \sum_{u_i, u_j \in \mathbf{F}_k} \|\hat{T}(u_i) - \hat{T}(u_j)\|_2^2 \\
& = \sum_{i,j} \mathbf{FI}(i,j) \|\mathbf{W}^\top \mathbf{FH}(:,i) - \mathbf{W}^\top \mathbf{FH}(:,j)\|_2^2 \\
& = \text{tr}(\mathbf{W}^\top \mathbf{FHL}_{\mathbf{FI}} \mathbf{H}^\top \mathbf{F}^\top \mathbf{W})
\end{aligned}$$

Then object function can be converted into:

$$\begin{aligned}
(3.14) \quad & \text{tr}(\mathbf{W}^\top \mathbf{XX}^\top \mathbf{W} - 2\mathbf{Y}^\top \mathbf{X}^\top \mathbf{W} + \mathbf{Y}^\top \mathbf{Y}) \\
& + \beta \text{tr}(\mathbf{W}^\top \mathbf{FHL}_{\mathbf{FI}} \mathbf{H}^\top \mathbf{F}^\top \mathbf{W}) \\
& = \text{tr}(\mathbf{W}^\top (\mathbf{XX}^\top + \beta \mathbf{FHL}_{\mathbf{FI}} \mathbf{H}^\top \mathbf{F}^\top) \mathbf{W} - 2\mathbf{Y}^\top \mathbf{X}^\top \mathbf{W}) \\
& = \text{tr}(\mathbf{W}^\top \mathbf{BW} - 2\mathbf{EW})
\end{aligned}$$

which completes the proof.

Following a similar approach to the CoFollowing relation, we can develop the relations of CoFollowed and Following below.

CoFollowed Relation: Let \mathbf{FE} be the CoFollowed matrix where $\mathbf{FE}(i,j) = 1$ if u_i and u_j are followed by at least one other person u_k . \mathbf{FE} can be obtained from the adjacency matrix \mathbf{S} , i.e., $\mathbf{FE} = \text{sign}(\mathbf{SS}^\top)$. Let $\mathbf{L}_{\mathbf{FE}}$ is the Laplacian matrix defined on \mathbf{FE} . Similarly, we can verify that the formulation for CoFollowed relation is equivalent to the following optimization problem:

$$(3.15) \quad \min_{\mathbf{W}} \text{tr}(\mathbf{W}^\top \mathbf{BW} - 2\mathbf{EW}) + \alpha \|\mathbf{W}\|_{2,1}$$

where \mathbf{B} and \mathbf{E} are defined as follows:

$$\begin{aligned}
(3.16) \quad & \mathbf{B} = \mathbf{XX}^\top + \beta \mathbf{FHL}_{\mathbf{FE}} \mathbf{H}^\top \mathbf{F}^\top \\
& \mathbf{E} = \mathbf{Y}^\top \mathbf{X}^\top
\end{aligned}$$

Following Relation: Let $\mathbf{L}_{\mathbf{S}}$ be the Laplacian matrix defined on \mathbf{S} . It is easy to verify that the formulation for Following relation is equivalent to the following optimization problem:

$$(3.17) \quad \min_{\mathbf{W}} \text{tr}(\mathbf{W}^\top \mathbf{BW} - 2\mathbf{EW}) + \alpha \|\mathbf{W}\|_{2,1}$$

where \mathbf{B} and \mathbf{E} are defined as follows:

$$\begin{aligned}
(3.18) \quad & \mathbf{B} = \mathbf{XX}^\top + \beta \mathbf{FHL}_{\mathbf{S}} \mathbf{H}^\top \mathbf{F}^\top \\
& \mathbf{E} = \mathbf{Y}^\top \mathbf{X}^\top
\end{aligned}$$

In this paper, we focus on the effect of each hypothesis on feature selection and do not consider the combination of multiple hypotheses into the same formulation to capture multi-faceted relations [23], which we leave as future work because it is more about how to use data

to effectively the values of different parameters. Closely examining the optimization problems for these four hypothesis, we can see that the LinkedFS framework is tantamount to solving the following optimization problem.

$$(3.19) \quad \min_{\mathbf{W}} \text{tr}(\mathbf{W}^\top \mathbf{BW} - 2\mathbf{EW}) + \alpha \|\mathbf{W}\|_{2,1}$$

Next we will present an optimization formulation to solve this problem and give convergence analysis.

3.3 An Optimal Solution for LinkedFS In this section, inspired by [19], we give a new approach to solve the optimization problem shown in Eq. (3.19). The Lagrangian function of the problem is:

$$(3.20) \quad \mathcal{L}(\mathbf{W}) = \text{tr}(\mathbf{W}^\top \mathbf{BW} - 2\mathbf{EW}) + \alpha \|\mathbf{W}\|_{2,1}$$

Taking the derivative of $\mathcal{L}(\mathbf{W})$,

$$(3.21) \quad \frac{\partial \mathcal{L}(\mathbf{W})}{\partial \mathbf{W}} = 2\mathbf{BW} - 2\mathbf{E}^\top + 2\alpha \mathbf{D}_{\mathbf{W}} \mathbf{W}$$

where $\mathbf{D}_{\mathbf{W}}$ is a diagonal matrix with the i -th diagonal element as⁶:

$$(3.22) \quad \mathbf{D}_{\mathbf{W}}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2}$$

All matrices \mathbf{B} defined above are semi-positive definite matrices and therefore $\mathbf{B} + \alpha \mathbf{D}_{\mathbf{W}}$ is positive definite matrix. Then setting the derivative to zero, we have:

$$(3.23) \quad \mathbf{W} = (\mathbf{B} + \alpha \mathbf{D}_{\mathbf{W}})^{-1} \mathbf{E}^\top$$

$\mathbf{D}_{\mathbf{W}}$ is dependent to \mathbf{W} and we proposed an iterative algorithm to obtain the solution \mathbf{W} . The detailed optimization method for LinkedFS is shown in Algorithm 1. We next verify that Algorithm 1 converges to the optimal \mathbf{W} , beginning with the following two lemmas.

LEMMA 3.1. *For any non-zero constants x and y , the following inequality holds [19].*

$$(3.24) \quad \sqrt{x} - \frac{x}{2\sqrt{y}} \leq \sqrt{y} - \frac{y}{2\sqrt{y}}$$

Proof. The detailed proof is similar as that in [19].

LEMMA 3.2. *The following inequality holds provided that $\mathbf{w}_t^i|_{i=1}^r$ are non-zero vectors, where r is an arbitrary*

⁶Theoretically, $\|\mathbf{W}(i,:)\|_2$ can be zero, then we can regularize $\mathbf{D}_{\mathbf{W}}(i,i)$ as $\mathbf{D}_{\mathbf{W}}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2 + \epsilon}$, where ϵ is a very small constant. It is easy to see that when $\epsilon \rightarrow 0$, then $\mathbf{D}_{\mathbf{W}}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2}$ approximates $\mathbf{D}_{\mathbf{W}}(i,i) = \frac{1}{2\|\mathbf{W}(i,:)\|_2}$

Algorithm 1 LinkedFS

Input: $\{\mathbf{F}, \mathbf{X}, \mathbf{Y}, \mathbf{S}, \mathbf{P}\}$ and the number of features expected to select, K ;

Output: K most relevant features

- 1: Construct \mathbf{E} and \mathbf{B} according to the hypothesis you choose;
 - 2: Set $t = 0$ and initialize $\mathbf{D}\mathbf{w}_t$ as an identity matrix;
 - 3: **while** Not convergent **do**
 - 4: Calculate $\mathbf{W}_{t+1} = (\mathbf{B} + \alpha\mathbf{D}\mathbf{w}_t)^{-1}\mathbf{E}^\top$;
 - 5: Update the diagonal matrix $\mathbf{D}\mathbf{w}_{t+1}$, where the i -th diagonal element is $\frac{1}{2\|\mathbf{w}_{t+1}(i,:)\|_2}$;
 - 6: $t = t + 1$;
 - 7: **end while**
 - 8: Sort each feature according to $\|\mathbf{W}(i,:)\|_2$ in descending order and select the top- K ranked ones.
-

number [19].

$$\begin{aligned}
& \sum_i \|\mathbf{w}_{t+1}^i\|_2 - \sum_i \frac{\|\mathbf{w}_{t+1}^i\|_2}{2\|\mathbf{w}_t^i\|_2} \\
(3.25) \quad & \leq \sum_i \|\mathbf{w}_t^i\|_2 - \sum_i \frac{\|\mathbf{w}_t^i\|_2^2}{2\|\mathbf{w}_t^i\|_2}
\end{aligned}$$

Proof. Substitute x and y in Eq. (3.24) by $\|\mathbf{w}_{t+1}^i\|_2$ and $\|\mathbf{w}_t^i\|_2^2$, respectively, we can see that the following inequality holds for any i .

$$(3.26) \quad \|\mathbf{w}_{t+1}^i\|_2 - \frac{\|\mathbf{w}_{t+1}^i\|_2}{2\|\mathbf{w}_t^i\|_2} \leq \|\mathbf{w}_t^i\|_2 - \frac{\|\mathbf{w}_t^i\|_2^2}{2\|\mathbf{w}_t^i\|_2}$$

Summing Eq. (3.26) over i , we see that Eq. (3.25) holds.

According to Lemma (3.2), we have the following theorem:

THEOREM 3.3. *At each iteration of Algorithm 1, the value of the objective function in Eq. (3.19) monotonically decreases.*

Proof. It can be easily verified that \mathbf{W}_{t+1} in line 4 of Algorithm 1 is the solution to the following problem,

$$\mathbf{W}_{t+1} = \arg \min_{\mathbf{W}} \text{tr}(\mathbf{W}^\top (\mathbf{B} + \alpha\mathbf{D}\mathbf{w}_t) \mathbf{W} - 2\mathbf{E}\mathbf{W}),$$

which indicates that,

$$\begin{aligned}
& \text{tr}(\mathbf{W}_{t+1}^\top (\mathbf{B} + \alpha\mathbf{D}\mathbf{w}_t) \mathbf{W}_{t+1} - 2\mathbf{E}\mathbf{W}_{t+1}) \\
& \leq \text{tr}(\mathbf{W}_t^\top (\mathbf{B} + \alpha\mathbf{D}\mathbf{w}_t) \mathbf{W}_t - 2\mathbf{E}\mathbf{W}_t)
\end{aligned}$$

That is to say,

$$\begin{aligned}
& \text{tr}(\mathbf{W}_{t+1}^\top \mathbf{B}\mathbf{W}_{t+1} - 2\mathbf{E}\mathbf{W}_{t+1}) + \alpha \sum_i \frac{\|\mathbf{W}_{t+1}(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2} \\
& \leq \text{tr}(\mathbf{W}_t^\top \mathbf{B}\mathbf{W}_t - 2\mathbf{E}\mathbf{W}_t) + \alpha \sum_i \frac{\|\mathbf{W}_t(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2}
\end{aligned}$$

Then we have the following inequality,

$$\begin{aligned}
& \text{tr}(\mathbf{W}_{t+1}^\top \mathbf{B}\mathbf{W}_{t+1} - 2\mathbf{E}\mathbf{W}_{t+1}) + \alpha \sum_i \|\mathbf{W}_{t+1}(i,:)\|_2 \\
& \quad - \alpha \left(\sum_i \|\mathbf{W}_{t+1}(i,:)\|_2 - \sum_i \frac{\|\mathbf{W}_{t+1}(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2} \right) \\
& \leq \text{tr}(\mathbf{W}_t^\top \mathbf{B}\mathbf{W}_t - 2\mathbf{E}\mathbf{W}_t) + \alpha \sum_i \|\mathbf{W}_t(i,:)\|_2 \\
& \quad - \alpha \left(\sum_i \|\mathbf{W}_t(i,:)\|_2 - \sum_i \frac{\|\mathbf{W}_t(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2} \right)
\end{aligned}$$

Meanwhile, according to Lemma 3.2, we have,

$$\begin{aligned}
& \sum_i \|\mathbf{W}_{t+1}(i,:)\|_2 - \sum_i \frac{\|\mathbf{W}_{t+1}(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2} \\
& \leq \sum_i \|\mathbf{W}_t(i,:)\|_2 - \sum_i \frac{\|\mathbf{W}_t(i,:)\|_2^2}{2\|\mathbf{W}_t(i,:)\|_2}
\end{aligned}$$

Therefore, we have the following inequality:

$$\begin{aligned}
& \text{tr}(\mathbf{W}_{t+1}^\top \mathbf{B}\mathbf{W}_{t+1} - 2\mathbf{E}\mathbf{W}_{t+1}) + \alpha \|\mathbf{W}_{t+1}\|_{2,1} \\
& \leq \text{tr}(\mathbf{W}_t^\top \mathbf{B}\mathbf{W}_t - 2\mathbf{E}\mathbf{W}_t) + \alpha \|\mathbf{W}_t\|_{2,1}
\end{aligned}$$

which indicates that the objective function of Eq. (3.19) monotonically decreases using the updating rules in Algorithm 1. Since $\mathbf{B} + \alpha\mathbf{D}\mathbf{w}_t$ is a positive definite matrix, the iterative approach in Algorithm 1 converges to an optimal solution, which completes the proof.

4 Experiments

In this section, we present the experiment details to verify the effectiveness of the proposed framework, LinkedFS. After introducing social media data used in experiments, we first confirm if linked data contains additional information than data randomly put together, then study how different relational information affects feature selection performance. In particular, we investigate how feature selection performance changes with different factors such as number of selected features, the amount of labeled data, which relational hypothesis impacts the performance most, and the relationships between the factors.

4.1 Social Media Data Two real-world social media datasets are Digg and BlogCatalog. For both datasets, we have posts and their social contextual information such as user-post relationship.

Digg: Digg⁷ is a popular social news aggregator that allows users to submit, digg and comment on stories. It also allows users to create social networks by

⁷<http://www.digg.com>

Table 1: Statistics of the Datasets

	BlogCatalog	Digg
# Posts	7,877	9,934
# Original Features	84,233	12,596
# Features after <i>TFIDF</i>	13,050	6,544
# Classes	14	15
# Users	2,242	2,561
# Following Relations	55,356	41,544
Ave # Posts	3.5134	3.8790
Max # Followers	820	472
Min # Followers	1	1
Network Density	0.0110	0.0063
Clustering Coefficient	0.3288	0.2461

designating other users as friends and tracking friends' activities. We obtain this dataset from [12]. The "following" relationships form a directed graph and the topics of stories are considered as the class labels.

BlogCatalog: BlogCatalog⁸ is a blog directory where users can register their blogs under predefined categories, which is used as class labels of blogs in our work. This dataset is obtained from [24]. The "following" relationships in BlogCatalog form an undirected graph, which means the CoFollowing and CoFollowed relationships in this dataset are the same.

The posts are preprocessed for stop-word removal and stemming. Obviously irrelevant features (terms) are also removed using *TFIDF*, a common practice in information retrieval and text mining [25]. Some statistics of these datasets are shown in Table 1.

4.2 Preliminary Verification Before conducting extensive experiments for feature selection, we validate if it is worthy of doing so. For the four relations, we form a null hypothesis for each: there is no difference between relational data and random data. If the null hypothesis is rejected, we then proceed to perform extensive experiments for feature selection. The difference is measured by a topic distance (T_{dist}) defined next.

Let \mathbf{c}_i be the class vector for post p_i , where $\mathbf{c}_i(j) = 1$ if p_i belongs to the class c_j , $\mathbf{c}_i(j) = 0$ otherwise. The topic distance between two posts, p_i and p_j , is defined as the distance between their class vectors, \mathbf{c}_i and \mathbf{c}_j :

$$(4.27) \quad T_{dist}^p(p_i, p_j) = \|\mathbf{c}_i - \mathbf{c}_j\|_2.$$

For each post p_i , we construct two vectors $\mathbf{cp}_t(i)$ and $\mathbf{cp}_r(i)$: the former by calculating the average T_{dist}^p between p_i and other posts from the same user, and the latter by calculating the average T_{dist}^p between p_i and randomly chosen posts from other users. The number of

Table 2: Statistics of T_{dist} to Support Relation Hypotheses ($\alpha = 0.01$)

	Digg	BlogCatalog
CoPost	<1.00e-14*	<1.00e-14*
CoFollowing	<1.00e-14*	2.80e-8*
CoFollowed	<1.00e-14*	1.23e-8*
Following	<1.00e-14*	1.23e-8*

randomly chosen posts is the same as the size of co-posts of p_i in the CoPost relation.

The topic distance between two users is defined as:

$$(4.28) \quad T_{dist}^u(u_i, u_j) = \|\bar{T}^*(u_i) - \bar{T}^*(u_j)\|_2;$$

where $\bar{T}^*(u_i) \in \mathbb{R}^k$, $\bar{T}^*(u_i) = \frac{\sum_{p_j \in \mathbf{F}_i} \mathbf{c}_j}{|\mathbf{F}_i|}$ is the topic interest distribution of u_i , and its j^{th} element represents the probability of u_i interested in the j^{th} class. In the same spirit of constructing $\{\mathbf{cp}_t, \mathbf{cp}_r\}$, for each user u_i , we construct $\{\mathbf{cf}_t(i), \mathbf{cf}_r(i)\}$, $\{\mathbf{cfe}_t(i), \mathbf{cfe}_r(i)\}$ and $\{\mathbf{fi}_t(i), \mathbf{fi}_r(i)\}$ by calculating their average $T_{dist}^u(u_i, u_j)$ according to CoFollowing, CoFollowed, and Following relations, respectively.

With the four pairs of vectors, we perform a two-sample t -test on each pair. The null hypothesis, H_0 , is that: there is no difference between the pair; the alternative hypothesis, H_1 , is that the average topic distance following a relation is less than that without. For example, for the CoPost relation, H_0 : $\mathbf{cp}_t = \mathbf{cp}_r$, and H_1 : $\mathbf{cp}_t < \mathbf{cp}_r$. The t -test results, p -values, are shown in Table 2⁹. The star (*) next to the p -value means that there is strong evidence ($p < 0.01$) to reject the null hypothesis. We observe that p -values for all four pairs are close to zero on both datasets. Hence, there is strong evidence to reject the null hypothesis. In other words, these relations are not random patterns and we now check how they help feature selection.

4.3 Quality of Selected Features and Determining Factors For both datasets, we randomly and evenly (50-50) split data into training data, \mathcal{T} and test data, \mathcal{U} . Following [19, 29, 26], feature quality is assessed via classification performance. If a feature subset is more relevant with the target concept, a classifier trained with the subset should achieve better accuracy [28]. Linear SVM [8] is used for classification. As a common practice, the parameters in feature selection algorithms and SVM are tuned via cross-validation. Since the performance of supervised

⁸<http://www.blogcatalog.com>

⁹We use the "ttest2" function from Matlab, which reports p -value as 0 if p -value is too small, i.e., exceeding the decimal places one allows. In our work, we use "<1.00e-14" when Matlab reports p -value as 0, which indicates it is significant.

learning improves with the number of labels data, we fix \mathcal{U} and sub-sample \mathcal{T} to generate training sets of different sizes, $\{\mathcal{T}_5, \mathcal{T}_{25}, \mathcal{T}_{50}, \mathcal{T}_{100}\}$, corresponding to $\{5\%, 25\%, 50\%, 100\%\}$ of \mathcal{T} , respectively. Another factor affecting learning performance is the number of features. Usually, other things being equal, the fewer features, the better. We set the numbers of selected features as $\{50, 100, 200, 300\}$.

Four representative feature selection algorithms are chosen as baseline methods: ttest (TT), Information Gain (IG), FishserScore (FS) [6], and Joint $\ell_{2,1}$ -Norms (RFS) [19]¹⁰ where RFS applies $\ell_{2,1}$ for both loss function and regularization, and FishserScore selects features by assigning similar values to the samples from the same class and different values to samples from different classes. We compare the four baseline methods with four methods based on LinkedFS, i.e., CoPost (CP), CoFollowing (CFI), CoFollowed (CFE), and Following (FI). The results are shown in Tables 3 and 4 for Digg and BlogCatalog, respectively. Since the Following relations in BlogCatalog are undirected, CFI is equivalent to CFE, having the same performance as shown in Table 4.

General trends. As seen in Tables 3 and 4, the performance of all the methods improves with increasing amount of labeled data. More often than not, with more features selected, the performance also improves. On both datasets, TT, IF, and FS perform comparably and RFS performs best. RFS selects features in batch mode and considers feature correlation. It is consistent what was suggested in [26, 29] that it is better to analyze instances and features jointly for feature selection.

Comparison with baselines. Our proposed methods, CP, CFI, CFE, and FI, consistently outperform all baseline methods on both datasets. Comparing with the best performance of baseline methods, the relative improvement of our methods is obtained and then averaged over different numbers of features. The results are given in Table 5. It is clear that CP and FI achieve better performance than CFI and CFE. That is, CoPost and Following hypotheses hold more strongly than CoFollowing and CoFollowed.

Relationships between amounts of labeled data and types of relations. Table 5 also says that our methods work more effectively when using small amounts of labeled data. For example, in Digg, CP is better than the best baseline by 14.54% with \mathcal{T}_5 , but only by 4.9% with \mathcal{T}_{100} . In Tables 3 and 4, if we select, for instance, 50 features, the performance using linked data with \mathcal{T}_5 is comparable with that without linked

Table 5: Classification Accuracy Improvement of the Proposed Methods

Improvement in Digg(%)				
Datasets	CP	CFI	CFE	FI
\mathcal{T}_5	+14.54	+7.01	+4.69	+15.25
\mathcal{T}_{25}	+4.59	+1.59	0	+4.02
\mathcal{T}_{50}	+7.19	+3.92	+1.05	+8.48
\mathcal{T}_{100}	+4.90	+3.15	+1.63	+4.64
Improvement in BlogCatalog(%)				
Datasets	CP	CFI	CFE	FI
\mathcal{T}_5	+10.71	+7.89	+7.89	+12.62
\mathcal{T}_{25}	+10.04	+5.00	+5.00	+9.50
\mathcal{T}_{50}	+9.70	+2.16	+2.16	+7.34
\mathcal{T}_{100}	+7.18	+0.46	+0.46	+7.67

with \mathcal{T}_{100} . In other words, linked data compensates the shortage of labeled data. The finding has its practical significance as in social media, it is not easy to obtain labeled data but there is often abundant linked data.

4.4 Effects of β and Numbers of Selected Features

An important parameter in LinkedFS is β that determines the impact of a relation on feature selection. A high value indicates the importance of this relation, or the corresponding hypothesis holds strongly. Another important parameter is the number of selected features. Hence, we study how the performance of CP, CFI, CFE, and FI varies with β and the number of selected features.

The results shown in Figures 3 and 4 are of \mathcal{T}_5 and \mathcal{T}_{50} of BlogCatalog data, respectively. Since CFI and CFE are equivalent for Blogcatalog, there are only three plots for CP, CFI, and CP. CP and FI achieve the peak performance with $\beta = 0.1$, and CFI with $\beta = 1e-6$. The performance patterns clearly vary with β and number of selected features. The results for Digg show similar patterns. CoPost and Following hypotheses hold more strongly than CoFollowing and CoFollowed hypotheses in the two datasets. Among the two parameters, performance is relatively more sensitive to the number of selected features. As pointed in [26], how to determine the number of selected features is still an open problem.

5 Related Work

Feature selection methods fall into three categories, i.e., the filter model, the wrapper model and embedded model [14]. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. Widely used filter-type feature selection methods include t-test, Information Gain, ReliefF and its multi-class extension

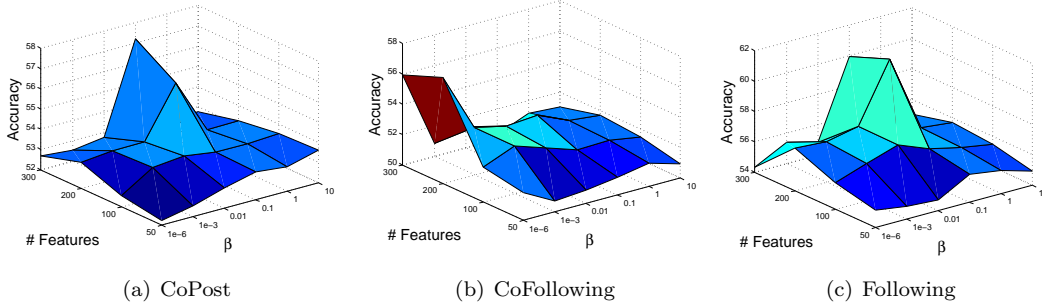
¹⁰We obtain the code for TT, IG and FS from featurselection.asu.edu, and RFS from the first author's webpage(sites.google.com/site/feipingnie)

Table 3: Classification Accuracy of Different Feature Selection Algorithms in Digg

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFI	CFE	FI
\mathcal{T}_5	50	45.45	44.50	46.33	45.27	58.82	54.52	52.41	58.71
	100	48.43	52.79	52.19	50.27	59.43	55.64	54.11	59.38
	200	53.50	53.37	54.14	57.51	62.36	59.27	58.67	63.32
	300	54.04	55.24	56.54	59.27	65.30	60.40	59.93	66.19
\mathcal{T}_{25}	50	49.91	50.08	51.54	56.02	58.90	57.76	57.01	58.90
	100	53.32	52.37	54.44	62.14	64.95	64.28	62.99	65.02
	200	59.97	57.37	60.07	64.36	67.33	65.54	63.86	67.30
	300	60.49	61.73	61.84	66.80	69.52	65.46	65.01	67.95
\mathcal{T}_{50}	50	50.95	51.06	53.88	58.08	59.24	59.39	56.94	60.77
	100	53.60	53.69	59.47	60.38	65.57	64.59	61.87	65.74
	200	59.59	57.78	63.60	66.42	70.58	68.96	67.99	71.32
	300	61.47	62.35	64.77	69.58	77.86	71.40	70.50	78.65
\mathcal{T}_{100}	50	51.74	56.06	55.94	58.08	61.51	60.77	59.62	60.97
	100	55.31	58.69	62.40	60.75	63.17	63.60	62.78	65.65
	200	60.49	62.78	65.18	66.87	69.75	67.40	67.00	67.31
	300	62.97	66.35	67.12	69.27	73.01	70.99	69.50	72.64

Table 4: Classification Accuracy of Different Feature Selection Algorithms in BlogCatalog

Datasets	# Features	Algorithms							
		TT	IG	FS	RFS	CP	CFE	CFI	FI
\mathcal{T}_5	50	46.54	40.96	41.31	46.16	53.37	53.01	53.01	52.84
	100	46.77	43.08	43.02	48.81	53.44	52.48	52.48	53.82
	200	46.84	44.06	45.66	50.77	55.94	53.61	53.61	57.30
	300	46.91	44.59	43.93	52.73	57.22	55.13	55.13	60.02
\mathcal{T}_{25}	50	48.13	40.58	45.44	47.60	53.40	53.24	53.24	52.79
	100	48.42	41.94	46.34	51.47	57.02	53.62	53.62	56.57
	200	48.05	43.45	53.07	53.64	58.83	55.81	55.81	60.50
	300	47.44	42.32	54.58	60.29	65.56	61.00	61.00	63.67
\mathcal{T}_{50}	50	48.66	52.21	48.23	52.51	56.22	53.47	53.47	56.97
	100	49.11	51.61	50.72	55.38	59.32	56.00	56.00	57.43
	200	48.43	51.54	53.74	62.02	68.08	63.58	63.58	65.66
	300	48.20	52.21	53.67	61.78	70.95	63.75	63.75	68.76
\mathcal{T}_{100}	50	50.54	54.33	52.39	54.55	58.34	55.31	55.31	55.92
	100	50.32	53.89	52.99	57.11	60.45	58.20	58.20	65.51
	200	50.77	54.02	54.80	66.33	70.81	63.11	63.11	68.31
	300	49.03	54.45	56.84	63.26	69.06	65.40	65.40	69.89

Figure 3: Performance Variation of Our Methods in \mathcal{T}_5 from BlogCatalog Dataset

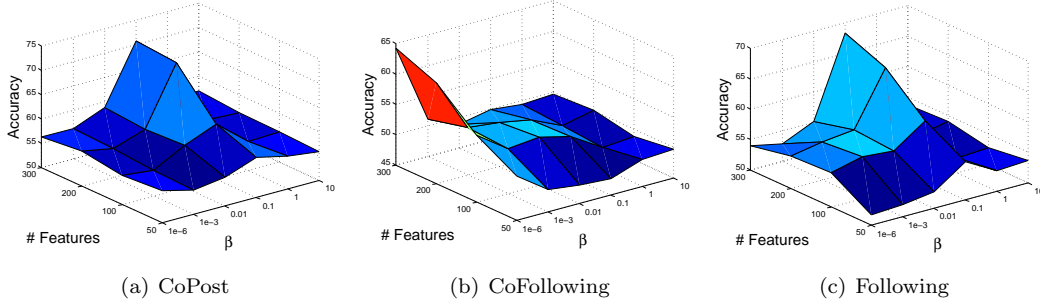


Figure 4: Performance Variation of Our Methods in \mathcal{T}_{50} from BlogCatalog Dataset

ReliefF [21], mRmR [20], LaplacianScore [10] and its extensions [28]. The wrapper model requires one pre-determined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance. Methods in that category include supervised algorithms [9, 6] and unsupervised algorithms [22, 7, 4]. However, these methods are usually computationally expensive [14] and may not be able to be applied on large scale data mining problems. For embedded model, the procedure of feature selection is embedded directly in the training process [2, 3].

Recently sparsity regularization such as $\ell_{2,1}$ of matrix in dimensionality reduction has been widely investigated and also applied into feature selection studies the $\ell_{2,1}$ of matrix is first introduced in [5] as rotational invariant ℓ_1 norm. A similar model for $\ell_{2,1}$ -norm regularization is proposed in [1, 15] to couple feature selection across tasks. Nie et al. [19] introduced a robust feature selection method emphasizing joint $\ell_{2,1}$ -norm minimization on both loss function and regularization. The $\ell_{2,1}$ -norm based loss function is robust to outliers in data points and $\ell_{2,1}$ -norm regularization selects features across all data points with joint sparsity. Zhao et al. [29] propose a spectral feature selection algorithm based on a sparse multi-output regression with a $\ell_{2,1}$ norm constraint, which can do well in both selecting relevant features and removing redundancy. Yang et al. [26] proposed a joint framework for unsupervised feature selection which incorporates discriminative analysis and $\ell_{2,1}$ -norm minimization. Different from existing unsupervised feature selection algorithms, this proposed algorithm selects the most discriminative feature subset from the whole feature set in batch mode.

The methods above focus on attribute-value data that is independent and identically distributed. There are recent developments that try to address relational data. In [11], the authors propose the problem of *relational feature selection*. Relational features are different from traditional features. A relational feature is, as an example in [11], $\text{Max}(\text{Age}(Y)) >$

65 where $\text{Movie}(x)$, $Y = \{y | \text{ActedIn}(x, y)\}$ where ActedIn is a relation that connects two objects x and y . Relational feature selection identifies a particular relation that links a single object to a set of other objects. Feature selection with linked data (or LinkedFS) still selects traditional features. Since LinkedFS involves more than one type (or source) of data such as user-post relationships and user-user relationships, it is related to *multi-source feature selection* (MSFS) [27] with the following differences: (1) sources in MSFS are different views of the same objects while additional sources in LinkedFS are different types of relations; and (2) MSFS and LinkedFS take different approaches to data of different sources: MSFS linearly combines multiple sources to a single source before applying single source feature selection, and LinkedFS considers a relation as a constraint.

6 Conclusions

Social media data differs from traditional data used in data mining. It presents new challenges to feature selection. In this work, we suggest to research a novel problem - feature selection with linked data. In particular, we extract four types of relations from linked data and propose a simple framework (LinkedFS) to integrate relational constraint into a state-of-the-art feature selection formulation. We further show that an optimal solution can be developed for LinkedFS, and conduct extensive experiments to show its efficacy and the relationships among several factors intrinsic to feature selection: numbers of selected features, percentages of labeled data, and importance of four types relations in performance improvement. This work aims to show the effectiveness of using linked data for feature selection. Our future work will focus on studying the combination of relations in a general model that can efficiently determine their contributions to feature selection, exploring additional and relevant information hidden in social media, and develop an open-source platform for collaborative research in this challenging new direction of feature selection.

Acknowledgments

The work is, in part, supported by an NSF grant(#0812551).

References

- [1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *NIPS*, 19:41, 2007.
- [2] G. Cawley and N. Talbot. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics*, 22(19):2348, 2006.
- [3] G. Cawley, N. Talbot, and M. Girolami. Sparse multinomial logistic regression via bayesian l1 regularisation. In *NIPS*, volume 19, page 209, 2006.
- [4] C. Constantinopoulos, M. Titsias, and A. Likas. Bayesian feature and model selection for gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1013–1018, 2006.
- [5] C. Ding, D. Zhou, X. He, and H. Zha. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd international conference on Machine learning*, pages 281–288. ACM, 2006.
- [6] R. Duda, P. Hart, D. Stork, et al. *Pattern classification*, volume 2. wiley New York, 2001.
- [7] J. Dy and C. Brodley. Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, 5:845–889, 2004.
- [8] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.
- [9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.
- [10] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *NIPS*, 18:507, 2006.
- [11] D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In *Proceedings of the 19th International Conference on Machine Learning*, pages 259–266. Citeseer, 2002.
- [12] Y. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher. Metafac: community discovery via relational hypergraph factorization. In *ACM SIGKDD*, pages 527–536. ACM, 2009.
- [13] H. Liu and H. Motoda. *Computational methods of feature selection*. Chapman & Hall, 2008.
- [14] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491, 2005.
- [15] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 339–348. AUAI Press, 2009.
- [16] P. Marsden and N. Friedkin. Network studies of social influence. *Sociological Methods and Research*, 22(1):127–151, 1993.
- [17] M. McPherson, L. S. Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.
- [18] S. Morris. Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology*, 56(12):1250–1273, 2005.
- [19] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint l21-norms minimization. *NIPS*, 2010.
- [20] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, pages 1226–1238, 2005.
- [21] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*, 53(1):23–69, 2003.
- [22] V. Roth and T. Lange. Feature selection in clustering problems. *NIPS*, 16:473–480, 2004.
- [23] J. Tang, H. Gao, and H. Liu. mtrust: Discerning multi-faceted trust in a connected world. In *the 5th ACM International Conference on Web Search and Data Mining*, 2012.
- [24] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering overlapping groups in social media. In *2010 IEEE International Conference on Data Mining*, pages 569–578. IEEE, 2010.
- [25] H. Wu, R. Luk, K. Wong, and K. Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):1–37, 2008.
- [26] Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou. L21-norm regularized discriminative feature selection for unsupervised learning. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
- [27] Z. Zhao and H. Liu. Multi-source feature selection via geometry-dependent covariance analysis. In *Journal of Machine Learning Research, Workshop and Conference Proceedings*, volume 4, pages 36–47. Citeseer.
- [28] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1151–1157. ACM, 2007.
- [29] Z. Zhao, L. Wang, and H. Liu. Efficient spectral feature selection with minimum redundancy. In *Proceedings of the Twenty-4th AAAI Conference on Artificial Intelligence (AAAI)*, 2010.