Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Exploring Social Context for Topic Identification in Short and Noisy Texts

Xin Wang[1,2,3], **Ying Wang**[1,2], **Wanli Zuo**[1,2*], **Guoyong Cai**[4]
[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, China
[3]School of Computer Technology and Engineering, Changchun Institute of Technology, Changchun 130012, China
[4]Guangxi Key Lab of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
xinwangjlu@gmail.com;zuowl@jlu.edu.cn

Abstract

With the pervasion of social media, topic identification in short texts attracts increasing attention in recent years. However, in nature the texts of social media are short and noisy, and the structures are sparse and dynamic, resulting in difficulty to identify topic categories exactly from online social media. Inspired by social science findings that preference consistency and social contagion are observed in social media, we investigate topic identification in short and noisy texts by exploring social context from the perspective of social sciences. In particular, we present a mathematical optimization formulation that incorporates the preference consistency and social contagion theories into a supervised learning method, and conduct feature selection to tackle short and noisy texts in social media, which result in a Sociological framework for Topic Identification (*STI*). Experimental results on real-world datasets from Twitter and Citation Network demonstrate the effectiveness of the proposed framework. Further experiments are conducted to understand the importance of social context in topic identification.

Introduction

Topic identification concerns a problem of predicting terms in text which indicate its subject or category, which has been extensively studied and applied in many content analysis applications, such as natural language processing (Murnane, Haslhofer, and Lagoze 2013; Vavliakis, Symeonidis, and Mitkas 2013), information retrieval (Jayarathna, Patra, and Shipman 2013), sentiment classification (Zheng et al. 2014; Li et al. 2014) and data mining (Ren and Wu 2013; Li, Jin, and Long 2012). With the pervasion of social media, short texts have become a popular mean of expression, through which users can easily produce content on various topics. Unlike a classical text with many words that provides sufficient word occurrences and helps gather sufficient statistics, a short text only consists of a few phrases or 1-2 sentences and may not carry enough contextual clues, even suffers from the severe issue of data sparsity; Hence, inferring unknown topics in short texts attracts increasing attention in recent years.

Existing topic identification algorithms can be roughly categorized into two groups: content-based approaches (Hu et al. 2009; Nicoletti, Schiaffino, and Godoy 2013; Chen et al. 2012), and content and link-based approaches (Weng et al. 2013; Vavliakis, Symeonidis, and Mitkas 2013; Takahashi, Tomioka, and Yamanishi 2014). The content-based approaches mainly fetch external text to expand the short text and then build a feature space directly for clustering and classification. For example, semantic expansion based on general knowledge (such as Open Directory Project(Bengel et al. 2004), Wordnet(Clifton, Cooley, and Rennie 2004) and Wikipedia (Egozi, Markovitch, and Gabrilovich 2011; Coursey, Mihalcea, and Moen 2009)) leverage the concepts or categories of knowledge sources to discover related topics. While the content and link-based approaches try to infer topics based on the combination of statistical topic modeling with network analysis. Many theoretical models and systems have been developed for topic identification, social network associated with short texts has rarely been studied.

In social sciences, it is well-perceived that user preferences or behaviors on specific topic categories play an important role in our social life and correlate with our social relations(Wang et al. 2014). The following two processes are proposed to explain social phenomena on popular topic categories: people are more likely to have consistent preferences in a specific short time interval, which has been recognized as preference consistency; people tend to influence others or their friends through a consequence of interactions and feedbacks based on specific topic categories, which has been recognized as social contagion. Inspired by these sociological observations, we explore the utilization of social context to facilitate topic identification in short and noisy texts.

In this paper, we aim to provide a supervised approach to topic identification in short and noisy texts by taking advantage of social context in tackling the noisy nature of messages. In particular, we first investigate whether social theories exist in the application of short and noisy texts. Then we discuss how the social context could be modeled and utilized for supervised topic identification. Finally, we conduct extensive experiments to verify the proposed model. Our contributions are summarized as follows,

- Demonstrate the existence of preference consistency theory and social contagion theory, and build message-message correlation relations;

- Propose a novel supervised framework, STI, to tackle short and noisy texts by integrating correlation relations between messages; and
- Evaluate STI on real-word datasets from Twitter and Citation Network, and elaborate the importance of different social theories on topic identification.

The rest of the paper is organized as follows. In Section 2, we analyze the real-world datasets and present motivating observations of social theories. In Section 3, we propose a novel framework for identifying topic categories. In Section 4, we report experimental results on real-world datasets with discussions. We finally conclude and present the future work in Section 5.

Notation

Let $T = [X, Y]$ be a corpus, where $X \in \mathbb{R}^{m \times n}$ is the message-feature matrix, $Y \in \mathbb{R}^{n \times c}$ is the topic label matrix, m is the number of features, n is the number of messages and c is the number of topic labels. For each message in the corpus $T = \{m_1, m_2, \cdots, m_n\}$, $m_i = (x_i, y_j) \in \mathbb{R}^{m+c}$ consists of short text messages and topic labels, where $x_i \in \mathbb{R}^m$ is the message feature vector and $y_i \in \mathbb{R}^c$ is the topic label vector. In this paper, we focus on several specific topic categories i.e., $c = 4$. It is practical to extend this setting to more topic categories. $\mathbf{u}=\{u_1, u_2, \cdots, u_d\}$, where d is the number of users in the corpus. $U \in \mathbb{R}^{d \times n}$ is a user-message matrix, where $U_{ij} = 1$ denotes that m_j is posted by user u_i. $R \in \mathbb{R}^{d \times d}$ is the user-user matrix, where $R_{ij} = 1$ indicates that user u_i is connected by user u_j.

We formally define topic identification of short and noisy text messages as,

Given a corpus of short text messages T with content X and corresponding topic labels Y, social relations for this corpus including the user-message relation U, and user-user relation R, we aim to learn a classifier W to automatically assign topic labels to unknown messages.

Data Analysis and Motivating Observations

In this section, we first collect two available datasets for this study, Obama-Romney Debate (ORD) from Twitter, and AMiner Citation Network Dataset (ACND) [1] from Citation Network respectively. Then, we present some motivating observations about social theories on topic identification.

Data Analysis

Both datasets consist of short text messages generated by users, social relations, and corresponding topic labels.

The first dataset is the Twitter data during the presidential debate in October 2012 between Barack Obama and Mitt Romney (ORD). We only extract four topics' debate: healthcare, immigration, foreign policy, and national defense authorization act (NDAA). The great majority of topic labels are annotated through hashtags in tweets, and only a quite small tweets without hashtags are manually labeled accrording to other tweets with hashtags at the same time interval. Social relations are built from retweet links and follow relations. All tweets whose authors have published fewer than two tweets are filtered.

The second dataset is the citation network (Tang et al. 2008) which contains the basic bibliographic information of computer science publications. We first extract all the papers published at four different conferences, VLDB, SIGIR, SIGKDD and NIPS. For each paper, we extract title and all authors. The title is taken as short text messages. Then, we construct social relations from the coauthor network. The authors with less than two papers published are filtered.

Some statistics of the datasets are summarized in Table1.

Table 1: Statistics of the Datasets

	ORD	ACND
# of Messages	12474	15996
# of Users	3712	7114
# of Relations	13872	35755
Avg. Messages per User	3.36	2.45
Avg. Lengths per Message	35.74	10.84

Motivating Observations of Social Theories about Popular Topic Categories

Recently, some researchers report the social theories related to topic categories in online social media, such as homophily (McPherson, Smith-Lovin, and Cook 2001), consistency (Abelson 1983) and contagion (Shalizi and Thomas 2011; Harrigan, Achananuparp, and Lim 2012). In this subsection, we investigate preference consistency and social contagion in popular topic categories via correlation between relational data and random data. Specifically, in the context of short texts, we pose two questions,

- Are the topic categories of two messages posted by the same user more likely to be consistent than those of two randomly selected messages?
- Are the topic categories of two messages posted by friends more likely to be similar than those of two randomly selected messages?

To answer the above two questions, we have to define the topic difference score as $D_{ij} = \parallel y_i - y_j \parallel$, where y_i denotes the topic label of message x_i. D_{ij} is 0, if y_i and y_j are the same topic label, otherwise 1. To answer the first question, we obtain two vectors $\mathbf{tv}_s$ and $\mathbf{tv}_r$ with equal number of elements. Each element of vector $\mathbf{tv}_s$ represents the topic difference score between m_i and m_j posted by the same user, while the element of vector $\mathbf{tv}_r$ denotes the topic difference score between m_i and another random m_r. Then, we construct a two-sample t-test on $\mathbf{tv}_s$ and $\mathbf{tv}_r$. The null hypothesis is H_0: $\mathbf{tv}_s = \mathbf{tv}_r$, is that there is no difference between the two vectors. The alternative hypothesis is H_1: $\mathbf{tv}_s < \mathbf{tv}_r$, is that the difference between messages with the same user is less than those messages with random users. For both datasets, the null hypothesis is rejected at significance level $\alpha = 0.01$ with p-value of 4.52e-20 and 2.47e-16 in ORD and ACND, respectively. The evidence from t-test suggests a positive answer to the first question: with high probability, the topic cat-

[1] http://arnetminer.org/AMinerNetwork

egories of two messages posted by the same user have higher consistency than those of two randomly selected messages.

For the second question, we also construct two vectors $\mathbf{iv}_f$ and $\mathbf{iv}_r$, where each element denotes the topic difference score between m_i and m_j posted by the users with friend relation. The null hypothesis is H_0: $\mathbf{tv}_f = \mathbf{tv}_r$ and the alternative hypothesis is H_1: $\mathbf{tv}_f < \mathbf{tv}_r$. The t-test results show that there is strong evidence (at significance level α = 0.01 with p-value of 5.51e-30 and 4.78e-23) to reject the null hypothesis in both tests on the above two datasets, which supports that the topic categories of two messages posted by friends are more likely to be similar than those of two randomly selected messages.

Positive answers to both questions provide evidence of the existence of preference consistency and social contagion in topic categories. With the verification of two problems, we next study how to exploit these two social theories for topic identification in short and noisy texts.

Our Framework

In this section, we first give the representation of short and noisy texts based on feature selection. Then, we model message content and correlation relations between messages based on social theories. Finally, we present our problem solution to implement topic identification in social media.

Introducing Feature Selection

Feature selection has recently emerged as a powerful means to obtain models of high-dimension data with high degree of interpretability, at low computational cost. A natural approach to topic identification is the lasso (Hastie et al. 2009), the penalization of the ℓ_1-norm of the estimator. The $\ell_{2,1}$-norm based linear reconstruction error minimization can lead to a sparse representation for the texts, which is robust to the noisy in features. The multi-class classifier can be learned by solving the following optimization problem,

$$\min_W \frac{1}{2} \parallel X^T W - Y \parallel_F^2 + \beta \parallel W \parallel_{2,1} \tag{1}$$

where W represents the learned classifiers, β denotes the sparse regularization parameter. In the objective function, the first term is Least Squares, which is employed to fit the learned model to message content. In terms of multi-class classification problems, the Least Squares aims to learn c classifiers by solving optimization problem. The second term is $\ell_{2,1}$-norm regularization on weight matrix W, which causes some of the coefficients to be exactly zero. The lasso can be regarded as a kind of continuous subset selection, which also controls the complexity of the model.

Representing Message Content

To find a better text representation for topic identification, we first perform a series of basic text preprocessing, such as removing stop-words and stemming, and then employ some feature selection models (Baccianella, Esuli, and Sebastiani 2013) to construct the text representation feature space. We investigate the following three widely used feature spaces for modeling message content,

- The unigram model (Unigram) with term presence, but not frequency;
- The term frequency model (TF); and
- The term frequency-inverse document frequency model (TF-IDF).

From these definitions of text representation, all feature values are normalized and ranged within [0, 1]. Note that in this paper, we do not consider the semantic annotation or expansion and leave it to our future work.

Modeling Social Context

We further extend framework to utilize social relations for topic identification. In order to transform user-message relations and social relations into correlation relations between messages, we model the following two social theories in topic identification,

- Preference consistency regularization: given user-message matrix U and user-user matrix R, the message-message correlation matrix for preference consistency ($\mathbf{IC}_{consistency}$) is defined as $\mathbf{IC}_{consistency} = U^T \times U$, where $\mathbf{IC}_{consistency} = 1$ indicates that m_i and m_j are posted by the same user, and topic categories of two messages are consistent. To integrate correlation relations between messages in topic identification, we build a latent connection to make two messages as close as possible if they are posted by the same user (*Consistency*). Under this scenario, it can be mathematically formulated as solving the following objective function,

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{IC}_{consistency}(ij) \parallel Y_{i*} - Y_{j*} \parallel^2 \\ &= \sum_{k=1}^{c} Y_{*k}^T(\mathbf{D} - \mathbf{IC}_{consistency})Y_{*k} \\ &= tr(Y^T \mathcal{L}_{consistency} Y) \end{aligned} \tag{2}$$

  where $tr(\cdot)$ denotes the trace of matrix. $Y = X^T W$ is the fitted value of topic label Y. $\mathcal{L}_{consistency} = \mathbf{D} - \mathbf{IC}_{consistency}$ is the Laplacian matrix (Chung 1997), where $\mathbf{IC}_{consistency} \in \mathbb{R}^{n \times n}$ is a message-message correlation matrix with preference consistency theory representing an undirected graph. $\mathbf{IC}_{consistency}(ij) = 1$ indicates that m_i is related to m_j, and $\mathbf{IC}_{consistency}ij = 0$ otherwise. $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{IC}_{consistency}(ij)$ indicating its diagonal element is the degree of a message in relation matrix $\mathbf{IC}_{consistency}$.
- Social contagion regularization: the message-message correlation matrix for social contagion ($\mathbf{IC}_{contagion}$) is defined as $\mathbf{IC}_{contagion} = U^T \times R \times U$, where $\mathbf{IC}_{contagion} = 1$ indicates that the author of m_i is a friend of the author who posts m_i, and topic categories of the two messages are the same. We build a latent connection to make two messages as close as possible if two users are retweet/coauthor with each other (*Contagion*); hence, it can be mathematically formulated as solving the following objective function,

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{IC}_{contagion}(ij) \parallel Y_{i*} - Y_{j*} \parallel^2 \\
&= \sum_{k=1}^{c} Y_{*k}^T(\mathbf{D}^{'} - \mathbf{IC}_{contagion})Y_{*k} \\
&= tr(Y^T\mathcal{L}_{contagion}Y)
\end{aligned} \tag{3}$$

where $\mathcal{L}_{contagion} = \mathbf{D}^{'} - \mathbf{IC}_{contagion}$ is the Laplacian matrix, where $\mathbf{IC}_{contagion} \in \mathbb{R}^{n\times n}$ is a message-message correlation matrix with social contagion theory to represent an undirected graph. $\mathbf{IC}_{contagion}(ij)$ = 1 indicates that m_i is related to m_j, and $\mathbf{IC}_{contagion}ij$ = 0 otherwise. $\mathbf{D}^{'} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $\mathbf{D}^{'}{}_{ii}$ = $\sum_{j=1}^{n}\mathbf{IC}_{contagion}(ij)$ indicating its diagonal element is the degree of a message in relation matrix $\mathbf{IC}_{contagion}$.

A Sociological Framework for Topic Identification

With the definition of preference consistency regularization and social contagion regularization, we propose a sociological framework for topic identification in short and noisy texts, STI, based on Laplacian regularization about message-message correlation relations while exploiting preference consistency theory and social contagion theory. STI is to solve the following optimization problem,

$$\begin{aligned}
&\min_{W\in Z} F(W) = \frac{1}{2} \parallel X^TW - Y \parallel_F^2 + \alpha tr(Y^T\mathcal{L}_{consistency}Y) \\
&+ \beta tr(Y^T\mathcal{L}_{contagion}Y) + \gamma \parallel W \parallel_{2,1} + \frac{\delta}{2} \parallel W \parallel_F^2
\end{aligned} \tag{4}$$

where $Z = \{W| \parallel W \parallel_{2,1} \leq z\}$, $z \geq 0$ is the radius of the ℓ_1-ball. α and β are positive regularization parameters, which are used to control the contributions of different correlation relations. γ is a positive parameter and denotes the sparse regularization parameter. δ is also a positive parameter and is used to prevent over-fitting. The last two terms are equivalent to elastic net (EN) regularization(Zou and Hastie 2005). By solving Eq.4, the topic label of each message can be predicted by,

$$arg\max_{i\in c} x^Tw_i \tag{5}$$

Next, we introduce an efficient algorithm to solve the optimization problem in Eq.4. We follow the formulation from Jun Liu et al. (Liu, Ji, and Ye 2009) using proximal gradient descent method, then W_{t+1} is updated in each step as,

$$W_{t+1} = arg\min_{W\in Z} G_{\lambda_t,W_t}(W) \tag{6}$$

where,

$$\begin{aligned}
&G_{\lambda_t,W_t}(W) = F(W_t) + \langle \bigtriangledown F(W_t), W - W_t\rangle \\
&+ \frac{\lambda_t}{2} \parallel W - W_t \parallel_F^2
\end{aligned} \tag{7}$$

Experiments

In this section, we conduct experiments to evaluate the effectiveness of our proposed framework STI. Through the experiments, we aim to answer the following two questions,

1. How effective is the proposed framework, STI, compared with other methods of topic classification?
2. What are the effects of topic correlation relations between messages on the performance of topic identification?

Experimental Settings and Evaluation Metric

The experimental settings of topic identification are described as follows: we randomly divide dataset A into two parts L and N. L possesses 80% of A used for training. The left 20% of A denoted as N is designated for testing. In each round of experiment, we choose x % of L as the amount of data used for training. x is varied as {10, 30, 50, 80, 100 }.

We follow the common metric for supervised learning to evaluate the performance of the proposed framework. In details, we take the labels of topic categories obtained by learning model as the set of identified results, denoting as R. Then, the Identification Accuracy (IA) is defined as,

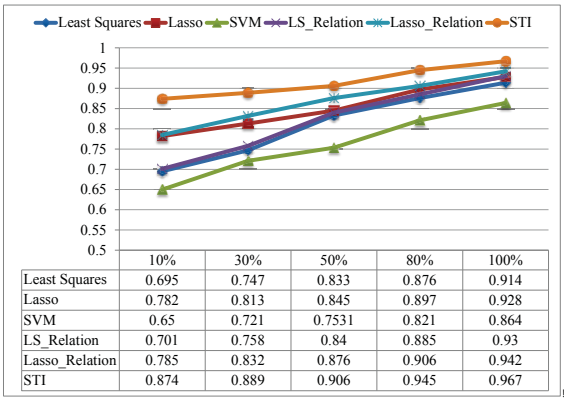
$$IA = \frac{|N \cap R|}{|N|} \tag{8}$$

where $|\cdot|$ denotes the size of a set. We use five-fold cross validation to ensure that our results are reliable, and average the result on the evaluation metric.

Performance Evaluation with Different Topic Classification Methods

To answer the first question, we compare the proposed framework, STI, with the following representative methods,

- *Least Squares*: Least squares (Hastie et al. 2009) is a widely used supervised classification method for independent and identically distributed data;
- *Lasso*: Lasso (Hastie et al. 2009) is one of the most popular feature selection methods, which is an alternative regularized version of least squares;
- *SVM*: Support Vector Machine (Suykens and Vandewalle 1999) is one of the most popular supervised learning models with associated learning algorithms that analyze data and recognize patterns, which solves classification and regression problems;
- *LS_Relation*: Least squares is applied on text content and topic correlation relations together. In this paper, the topic correlation relations are used as feature expansion for the feature space of each short text. If a short text m_i is correlated to another short text m_j, we add the features of short text m_j into m_i's feature vector. The following $Lasso\_Relation$ methods use the same method to expand feature space; and
- *Lasso_Relation*: Lasso is applied on text content and topic correlation relations together, which is the sparse version of the method of $LS\_Relation$.



	10%	30%	50%	80%	100%
Least Squares	0.695	0.747	0.833	0.876	0.914
Lasso	0.782	0.813	0.845	0.897	0.928
SVM	0.65	0.721	0.7531	0.821	0.864
LS_Relation	0.701	0.758	0.84	0.885	0.93
Lasso_Relation	0.785	0.832	0.876	0.906	0.942
STI	0.874	0.889	0.906	0.945	0.967

(a) ORD on Twitter

	10%	30%	50%	80%	100%
Least Squares	0.529	0.689	0.795	0.818	0.851
Lasso	0.715	0.791	0.841	0.893	0.908
SVM	0.697	0.751	0.7981	0.842	0.895
LS_Relation	0.672	0.705	0.821	0.874	0.88
Lasso_Relation	0.731	0.815	0.872	0.901	0.917
STI	0.826	0.863	0.896	0.912	0.941

(b) ACND on Citation Network

Figure 1: Topic Identification Accuracy on ORD and ACND

In this group of experiments, we first set $\alpha = \beta$ which means that the two correlation relation matrices are simply combined with equal weight. Then, we tune γ and δ via common cross-validation. For general experiment purposes, we set $\alpha = \beta = 0.1$ for ORD, and $\alpha = \beta = 0.5$ for ACND. Let set $\gamma = 0.5$ for ORD and $\gamma = 0.1$ for ACND experimentally. Similarly, we set $\delta = 1$ for ORD and $\delta = 0.2$ for ACND experimentally. The experiment results are shown in Figure 1.

By comparing the results of different methods, we draw the following observations,

- Our proposed framework consistently outperforms other baseline methods based on the same performance metric, namely, STI achieves better performance than the other baseline methods when using a small training dataset;
- Compared with the text-based methods, the performance of $LS\_Relation$, $Lasso\_Relation$ and STI is better than $Least\ Squares$, $Lasso$ and SVM, which indicates that these integrated relations methods are able to achieve better performance by introducing social relations into feature augmentation of the text feature space and regularization. In particular, our framework obtains significant improvement;
- The feature selection-based methods, such as $Lasso$, $Lasso\_Relation$ and STI, achieve better performance than $Least\ Squares$ and SVM, which suggest that a sparse solution of the feature space is an effective way to tackle with short and noisy texts. The introduction of sparse regularization has positive impacts on the proposed topic identification method; and
- Among the methods of incorporating social relations, STI obtain the best performance of both datasets with different sizes of training data, which demonstrates that the regularization way of correlation relations between messages outperforms the text feature space augmentation based on social relations.

In summary, we perform t-test on all comparisons and the t-test results suggest that all improvement is significant. With the help of regularization based on correlation relations between messages, the proposed framework, STI, gains significant improvement over representative baseline methods, which answers the first question asked at the beginning of this section that STI is effective in topic identification compared with other baseline methods.

Impact of Correlation Relations between Messages in STI

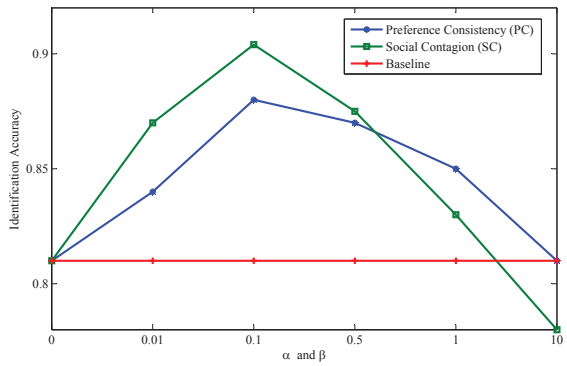
In this subsection, we study the importance of topic correlation relations between messages in STI and accordingly answer the second question at the beginning of this section.

We conduct experiments with seperating two topic correlation relations to further understand the impact of each relation on the performance of topic identification. We take STI without correlation relations as the baseline method. In this experiment, the value of α and β are varied as $\{0, 0.01, 0.1, 0.5, 1, 10\}$, and the results are shown in Figure 2 for ORD and ACND, respectively. "PC" and "SC" denote that the performance of preference consistency and social contagion, respectively.

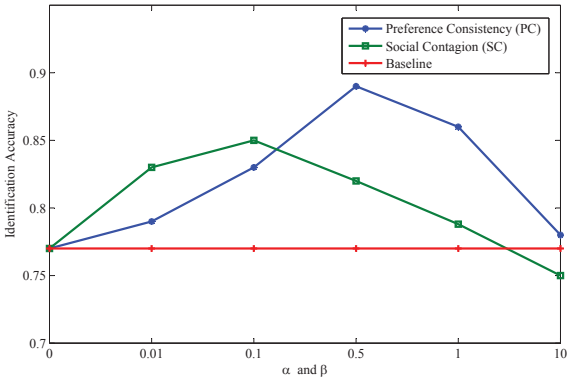
By comparing the results of different α and β, we draw the following observations,

- The results suggest that the proposed framework STI can achieve relatively good performance when α and β are in the range of [0.1, 0.5]. The curves of PC reach the peak at $\alpha = \beta = 0.1$, and the curves of SC reach the peak at $\beta = 0.1$ and $\beta = 0.5$ respectively;
- Intuitively, for most tuned parameters, the identification accuracy with PC and SC is higher than the baseline method, which suggests that PC and SC are not sensitive to parameter setting. Hence, it is not necessary to make much effort to tune the parameters. In addition, with any one of topic correlation relations, STI can also improve the performance of topic identification as well; and
- With different parameter settings, SC has a stronger impact on the framework than PC on ORD dataset, conversely, PC has a stronger impact than SC on ACND dataset. A potential reason is that the types of social relations are different on two datasets.

In summary, the results of experiments further demonstrate the importance of modeling topic correlation relations



(a) ORD on Twitter



(b) ACND on Citation Network

Figure 2: Performance Variation of STI

between messages in topic identification, which correspondingly answer the second question. In addition, an appropriate combination of feature selection model and relation regularizations based on social theories can greatly improve the performance of topic identification.

Impact of Data Representation Methods for Text Content

There are many measurements for feature selection and in this subsection, we investigate the impact of different feature selection measurements in the proposed framework STI. In this experiment, we fix the number of features as 3000. Table 2 demonstrates the performance of the proposed framework with different feature selection measurements on ACND. We omit the results on ORD since we have similar observations. From Table 2, we observe that,

Table 2: Different Measurements of Feature Selection for Text Representation

x	Unigram	TF	TF-IDF
10%	0.826	0.867	0.874
30%	0.842	0.87	0.889
50%	0.855	0.884	0.906
80%	0.89	0.923	0.945
100%	0.917	0.941	0.967

- Different measurements may lead to different performance for STI. The performance of "TF" and "TF-IDF" is very similar, which are much better than that of "Unigram"; and
- Feature selection for text representation also plays an important role in the supervised learning model.

The above results show the importance of feature selection for text representation in topic identification by modeling message-feature matrix in our framework.

Conclusion and Future Work

With the pervasion of social media, short texts have become a popular mean of expression, through which users can easily produce content on various topics. As the classical machine learning approaches heavily rely on the term co-occurrence information, the short and noisy texts unduly influence the significant improvement of the performance about topic identification. In this paper, we investigate topic identification in short and noisy texts by exploring social context from the perspective of social sciences, which results in a framework, STI. In particular, we first verify existence of social theories in short texts. Then we model preference consistency theory and social contagion theory in message-message correlation relations. Finally, a sociological framework is proposed for identifying topic categories. Experimental results on a real-world datasets from Twitter and Citation Network demonstrate the effectiveness of the proposed framework.

There are several interesting directions for future work. It is interesting to explore whether: (1) we plan to consider the semantic annotation and expansion based on external resources such as Wikipedia and WordNet for improving the accuracy of topic identification in short and noisy texts; (2) we will explore social theories and social context on other applications with short and noisy texts, such as sentiment classification and word sense disambiguation.

Acknowledgments

We truly thank the anonymous reviewers for their pertinent comments. In addition, we truly thank the help of DMML at ASU. This work is supported by the National Natural Science Foundation of China under grant No.61300148; the scientific and technological break-through program of Jilin Province under grant No.20130206051GX; the Science Foundation for China Postdoctoral under grant No.2012M510879; the science and technology development program of Jilin Province under grant No.20130522112JH; Guangxi Key Lab of Trusted Software under grant No.kx201420.

References

Abelson, R. P. 1983. Whatever became of consistency theory? *Personality and Social Psychology Bulletin*.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2013. Using micro-documents for feature selection: The case of ordinal text classification. *Expert Systems with Applications* 40(11):4687–4696.

Bengel, J.; Gauch, S.; Mittur, E.; and Vijayaraghavan, R. 2004. Chattrack: Chat room topic detection using classification. In *Intelligence and Security Informatics*. Springer. 266–277.

Chen, Y.; Li, Z.; Nie, L.; Hu, X.; Wang, X.; Chua, T.-s.; and Zhang, X. 2012. A semi-supervised bayesian network model for microblog topic classification. In *COLING*, 561–576.

Chung, F. R. 1997. *Spectral graph theory*, volume 92. American Mathematical Soc.

Clifton, C.; Cooley, R.; and Rennie, J. 2004. Topcat: data mining for topic identification in a text corpus. *Knowledge and Data Engineering, IEEE Transactions on* 16(8):949–964.

Coursey, K.; Mihalcea, R.; and Moen, W. 2009. Using encyclopedic knowledge for automatic topic identification. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 210–218. Association for Computational Linguistics.

Egozi, O.; Markovitch, S.; and Gabrilovich, E. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)* 29(2):8.

Harrigan, N.; Achananuparp, P.; and Lim, E.-P. 2012. Influentials, novelty, and social contagion: The viral power of average friends, close communities, and old news. *Social Networks* 34(4):470–480.

Hastie, T.; Tibshirani, R.; Friedman, J.; Hastie, T.; Friedman, J.; and Tibshirani, R. 2009. *The elements of statistical learning*, volume 2. Springer.

Hu, X.; Sun, N.; Zhang, C.; and Chua, T.-S. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 919–928. ACM.

Jayarathna, S.; Patra, A.; and Shipman, F. 2013. Mining user interest from search tasks and annotations. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 1849–1852. ACM.

Li, F.; Wang, S.; Liu, S.; and Zhang, M. 2014. Suit: A supervised user-item based topic model for sentiment analysis. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Li, L.; Jin, X.; and Long, M. 2012. Topic correlation analysis for cross-domain text classification. In *AAAI*.

Liu, J.; Ji, S.; and Ye, J. 2009. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 339–348. AUAI Press.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 415–444.

Murnane, E. L.; Haslhofer, B.; and Lagoze, C. 2013. Reslve: leveraging user interest to improve entity disambiguation on short text. In *Proceedings of the 22nd international conference on World Wide Web companion*, 1275–1284. International World Wide Web Conferences Steering Committee.

Nicoletti, M.; Schiaffino, S.; and Godoy, D. 2013. Mining interests for user profiling in electronic conversations. *Expert Systems with Applications* 40(2):638–645.

Ren, F., and Wu, Y. 2013. Predicting user-topic opinions in twitter with social and topical context.

Shalizi, C. R., and Thomas, A. C. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* 40(2):211–239.

Suykens, J. A., and Vandewalle, J. 1999. Least squares support vector machine classifiers. *Neural processing letters* 9(3):293–300.

Takahashi, T.; Tomioka, R.; and Yamanishi, K. 2014. Discovering emerging topics in social streams via link-anomaly detection. *Knowledge and Data Engineering, IEEE Transactions on* 26(1):120–130.

Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su, Z. 2008. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, 990–998.

Vavliakis, K. N.; Symeonidis, A. L.; and Mitkas, P. A. 2013. Event identification in web social media through named entity recognition and topic modeling. *Data & Knowledge Engineering* 88:1–24.

Wang, S.; Hu, X.; Yu, P. S.; and Li, Z. 2014. Mmrate: inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1246–1255. ACM.

Weng, L.; Ratkiewicz, J.; Perra, N.; Gonçalves, B.; Castillo, C.; Bonchi, F.; Schifanella, R.; Menczer, F.; and Flammini, A. 2013. The role of information diffusion in the evolution of social networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 356–364. ACM.

Zheng, X.; Lin, Z.; Wang, X.; Lin, K.-J.; and Song, M. 2014. Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. *Knowledge-Based Systems* 61:29–47.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.