doi:10.3969/j.issn.1007-7162.2016.06.015

基于混合高斯分布伪样本生成的情感分析方法

梁礼欣，郝志峰，蔡瑞初，温　雯
（广东工业大学 计算机学院，广东 广州 510006）

摘要：针对微博行文自由性大，情感倾向识别困难的问题，提出了一种基于混合高斯分布伪样本生成技术和条件随机场模型的新方法。该方法首先利用混合高斯分布模型来为训练集中的少数类生成伪样本从而构建一个情感倾向分布平衡的训练集，然后通过使用 Word2vec 来扩展微博句子以丰富它的情感信息，从而缓解情感词典不足够大对情感分类的负面影响；最后将条件随机场模型应用在上面已经平衡和扩展后的训练集上. 实验结果表明该方法比现有方法在数据集情感倾向分布不平衡时能更有效地识别微博的情感倾向.

关键词：情感分析；混合高斯分布；条件随机场；情感倾向；不平衡性；Word2vec
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1007-7162(2016)06-0085-06

An Approach to Sentiment Analysis of Chinese Microblogs Based on Gaussian Mixture Distribution Pseudo-sample Generation

Liang Li-xin, Hao Zhi-feng, Cai Rui-chu, Wen Wen
(School of Computers, Guangdong University of Technology, Guangzhou, 510006)

Abstract: Since informal words and expressions are widely used in miscroblogs, sentiment analysis of the microblogs is a difficult scientific problem, especially with the data in imbalanced sentiment distribution. GWCRF (Gaussian Mixture Distribution Word2vec CRF), a method based on pseudo-sample generation technique and Conditional Random Field (CRF) for sentiment analysis of microblogs in imbalance distribution is presented. In the proposed method, firstly, the Gaussian Mixture Distribution is leveraged to generate pseudo-samples, which can increase the samples of minor classes for balancing the train data sets. Secondly, Word2vec technology is leveraged to enrich the microblog message and overcome the problem that sentiment lexicon is not large enough. Moveover, the CRF model is proposed to apply in the above balanced and extended train data sets. Experimental results on the microblog data demonstrate that this method outperforms the state-of-art methods in sentiment analysis of the microblog data sets with imbalanced sentiment distribution.

Key words: sentiment analysis; Gaussian mixture distribution; conditional random field; sentiment; imbalance; Word2vec

微博作为一个新的社交平台，承载了海量的信息，如何有效分析和挖掘用户微博中的情感是非常有意义的[1]. 与传统的情感分析工作一样，对微博的情感分析方法可以分为两类. 一类是基于情感词典和规则的方法，这类方法通过计算句子中负面情感词和正面情感词的个数来识别情感倾向[2-6]. 另一类是基于机器学习的方法，它们通过挑选合适的特征来训练模型[7-11].

然而，传统方法都没有意识到中文微博数据集中的情感倾向分布不平衡性对情感分类的影响. 人们在微博中讨论的话题往往带有很强的情感倾向性，这导致很多话题的情感倾向分布不平衡，例如“#90 后暴打老人#”等话题本身具有明显的贬义情感，而“#莫言获诺贝尔奖#”这个话题具有明显的褒

收稿日期：2016-03-23
基金项目：国家自然科学基金资助项目(61472089,61572143)
作者简介：梁礼欣(1990-)，男，硕士研究生，主要研究方向为文本情感分析、数据挖掘.

义情感[12].

数据集情感倾向分布的不平衡性恰恰是导致很多机器学习算法表现不好的重要因素,尤其是在情感倾向中占少数的类别的识别效果上[2].此外,微博的长度比传统文本要短,这使得传统方法很难从其中抽取出很多有助于情感分类的信息,而且目前还没有一个足够大的情感词典可以覆盖所有的情感词.针对以上问题,本文提出了一种基于混合高斯分布伪样本生成技术和条件随机场(Conditional Random Field,CRF)模型的方法 GWCRF(Gaussian Mixture Distribution Word2vec CRF).实验结果表明,在中文微博情感倾向分析任务上,GWCRF 方法比现有的方法取得更好的效果.

1 相关工作

这个章节将介绍中文微博情感分析任务的相关研究成果. Barbosa[8]的情感分析方法首先是将推特分为主观句或者客观句,然后判别主观句的情感倾向是正面或者是负面. Davidov[13]使用一个 KNN 相似的分类器去对推特进行情感倾向分析,该方法将推特的表情符号和 hashtags 主题标签作为特征. Vanzo[9]提出应用 SVMhmm 算法在包含上下文信息(例如微博的主题和对原始微博的回复等)的推特上. Jiang[14]也提出了一种与文献[8]中的方法不同的两步的情感分类方法,它考虑了目标依赖特征和基于图的情感优化.

虽然英文微博的情感分析已经有很多成果,但是中文微博情感分析还处于起步阶段.中文与英文不同的是它具有更加复杂的句子类型和语言结构.这导致了很多英文微博情感分析方法在一定程度上不适合使用在中文上.

Xie[1]首次尝试对句子级别的中文微博进行情感分类,比较了三种方法的效果:基于情感词典的方法、基于表情符合的方法和使用 SVM 算法的分层混合方法.最后,实验结果显示混合方法达到最好的效果.该方法首先把一条微博分为几个句子,接着使用已经训练好的 SVM 分类器来判别每个句子的情感倾向.与文献[8]中的方法不同,SVM 分类器一步就将句子分为正面、负面或者中性,这样做的效果比两步的方法更好.但是他们并没有考虑这些句子间的依赖关系.另外,现有的方法都没有意识到数据集中的情感倾向分布不平衡性对情感倾向分析的影响,从而导致分类器偏向于多数类样本,使得少数类样本识别性能不高.

2 中文微博情感分析方法

2.1 利用混合高斯分布生成伪样本

高斯混合模型是用高斯概率密度函数精确地量化事物,它是一个将事物分解为若干的基于高斯概率密度函数形成的模型.

中文微博数据集中情感倾向分布是不规则的,不能以一种单一的分布函数对软件缺陷分布进行模拟.传统方法以每个样本为中心独立添加伪样本,这样不仅不能很好地刻画样本分布,而且容易导致样本重叠.另外,用基于样本全局的分布的单高斯模型来生成伪样本会破坏样本原先的分布.基于上面的分析,本文提出利用混合高斯分布生成伪样本,具体步骤如下:

(1) 对于一个训练集 t_1,将它分为多数类 maj_1(即数据集中情感倾向占多数的类别)和少数类 min_1(即数据集的情感倾向中占少数类别).例如,在不平衡数据集"#90 后暴打老人#"中,负面情感的数据会比正面情感数据的多很多,所以负面情感的数据就是多数类,正面情感的数据就是少数类.

(2) 对于少数类 min_1,使用 Affinity Propagation 聚类算法将它聚成 m 个子类,假设 $\{X_i\}$ 代表 min_1,那么聚类后的 min_1 就可以表示为 $\{X_i^1, X_i^2, \cdots, X_i^m\}$,$\{X_i^j\}$ 代表 min_1 中第 j 个子类.

(3) 为了构建平衡的数据集,利用混合高斯分析按比例为 min_1 中的每个子类随机生成伪样本得到 min_2,使得 min_2 的样本数量与 min_1 的样本数量相近.为每个子类进行高斯参数估计,N_j 是指 min_1 中第 j 个子类 $\{X_i^j\}$ 的样本数,具体步骤如下:

① 计算 min_1 中第 j 个子类 $\{X_i^j\}$ 的均值

$$\mu^j = \frac{1}{N_j}\sum_{i=1}^{N_j} X_i^j ; \tag{1}$$

② 计算子类 $\{X_i^j\}$ 的协方差矩阵 $\boldsymbol{U}$;

③ 根据协方差矩阵 $\boldsymbol{U}$ 和均值 μ^j,为子类 $\{X_i^j\}$ 生成符合高斯分布的伪样本;

④ 将数据集 min_2 和 maj_1 集中在一起得到一个平衡训练集 t_2.然后将 t_2 代替 t_1 作为最终的训练集.

2.2 利用 Word2vec 来扩展微博

Word2vec 是 Google 在 2013 年中开源的一款将词表征为实数值向量的高效工具,其利用深度学习的思想,把对文本内容的处理简化为 K 维向量空间中的向量运算,而向量空间上的相似度可以用来表示文本语义上的相似度.

本文从新浪微博 API 收集了大量的微博语料来训练词向量,对微博进行清洗过滤后,剩下 6 G 的微博数据作为训练集. 接着使用 Word2vec 中的 Skip-gram 模型来训练词向量,这样就可以通过该词向量来求微博中每个词的相似词了.

本文扩展微博方法的步骤如下:(1)对于一条微博 t,将它分词之后得到它的词序列,表示为(W_1, W_2,…,W_n). (2)使用已经训练好的词向量来求微博 t 中每个词的前 k 个相似词,从而达到扩展微博的目的. 扩展后的微博可以表示为(W_1,W_2,…,W_n, W_{11},W_{12},…,W_{1k},W_{21},W_{22},…,W_{2k},…,W_{n1},W_{n2},…, W_{nk}),其中(W_{11},W_{12},…,W_{1k})代表词 W_1 的前 k 个相似词. (3)对于微博中表情符号和标点符号的处理是将它们直接保留在微博中,所以扩展后的微博会比原微博含有更多的信息.

2.3　条件随机场模型

CRF 模型是由 Lafferty 在 2001 年提出的一种典型的判别式模型. CRF 模型不仅拥有判别式模型的优点,而且拥有产生式模型考虑到上下文标记间的转移概率,以序列化形式进行全局参数优化和解码的特点. 它还解决了其他判别式模型难以避免的标记偏置问题. 在 CRFs 模型中,应用最广泛的是 Linear-chain CRF 模型,下面介绍怎样将它应用在中文微博情感分析问题上.

在图 1 中,X 是一个观测变量集合,例如一个样本序列,每个变量 X 代表情感分析任务中的一个样本,Y 是要预测的目标变量的集合,例如在情感分析任务中,Y 代表样本的标签,取值范围是{正面,负面}.

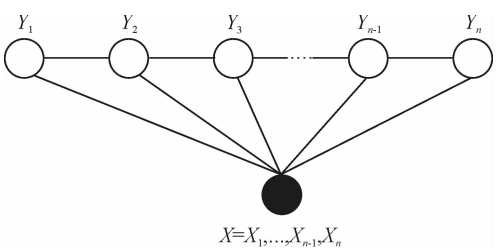

图 1　线性链条件随机场

Fig. 1　Graphical representation of Linear-Chain CRF

CRF 模型公式化表示为

$$P(Y|X)=\frac{1}{Z(X)}\exp\{\sum_{k=1}^{T}\lambda_k f_k(y,x)\} \qquad (2)$$

和

$$Z(X)=\sum_{Y}\exp\{\sum_{k=1}^{T}\lambda_k f_k(y,x)\}, \qquad (3)$$

这里$\{f_k\}$是特征集合,包括状态特征和转移特征,$\{\lambda_k\}$是特征权重的集合. $\{f_k\}$可以表示为

$$f_k(y,x)=\sum_{i=1}^{n} f_k(y_{i-1},y_i,x,i),k=1,2\cdots,T. \qquad (4)$$

Linear-chain CRF 模型的定义表明,每个特征函数都可以依赖任何时间点的观测变量. 在图 1 中,$X=(X_1,X_2,\cdots,X_n)$作为一个单独的观测变量节点,而不是将每个变量 $X_1,X_2,\cdots,X_n$ 用 L-BFGS 拟牛顿法来估计模型的参数,利用 Loopy BP(Loopy Belief Propagation)[15]算法来推理测试数据的标签序列.

2.4　中文微博情感分析的流程

中文微博情感分析任务可以视为序列标注任务,目标是为每个样本打上 Y 或 N 的标签,Y 是指正面情感,N 是指负面情感.

图 2 是 GWCRF 方法的工作流程,具体细节如下.

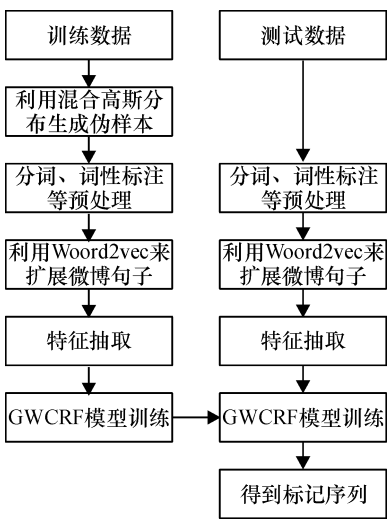

图 2　中文微博情感分析流程图

Fig. 2　Workflow of Chinese microblog sentiment analysis

(1) 利用混合高斯分布生成伪样本:为训练集中的少数类的每一个子类生成高斯随机伪样本.

(2) 预处理:本文实验中采用开源的中科院 ICTCLAS 中文分词器对原始的微博数据集进行分词和词性标注,同时它允许用户添加自定义的分词词典. 由于微博语料中情感表达的方式十分口语化,因此本文通过构建自己的分词词典来提高分词效果,分词词典主要由情感词词典、网络用语词典等构成. 网络用语词典不仅包括网络流行词或短语如:屌丝、碉堡等,还包括一些表达观点的日常口语词如:傻逼、屁用等.

(3) 利用 Word2vec 扩展微博:使用 Skip-gram 模型训练的词向量来扩展微博句子.

(4) 特征抽取:通过对比多组特征组合的实验效果,本文最终确定了一组最优特征组合,所采用的

最优特征组合特征如表1所示.

(5) 训练GWCRF模型:输入为训练集中的特征向量,使用L-BFGS算法来估计每个特征的权重.

(6) 得到标记序列:输入一个测试集的特征向量到已经训练好的模型得到标签序列.

表1 最优特征组合的具体情况

Tab.1 The details of the best feature combination

特征名	特征描述
has_positive_emoticons	是否含有正面表情符号
has_negative_emoticons	是否含有负面表情符号
positive_emoticons_more	正面表情符号是否比负面表情符号多
has_positive_buzzwords	是否含有正面网络流行词
has_negative_buzzwords	是否含有负面网络流行词
positive_buzzwords_more	正面网络流行词是否比负面网络流行词多
word_feature	过滤停用词后句子中所剩词的数量
rate_of_positive_words	正面情感词占句中总词数的比例
has_hyperlink	句中是否含有超链接
positive_words_more	正面情感词是否比负面情感词多
has_adj	是否含有形容词
has_ single_? /!	是否含有单个? 或!
rate_of_negative_words	负面情感词占句中总词数的比例
sentiment_phrase	是否含有情感短语,例如没文化、没意思等
has_ continuous_? /!	是否含有连续的? 或!,例如????、!!!!! 等

3 实验结果与分析

3.1 实验设置

本文通过三个实验来验证GWCRF方法的有效性.(1) 比较GCRF方法(即将CRF模型直接应用在经过混合高斯分布平衡处理过的训练集上)和CRF方法(即将CRF模型直接应用在没有经过平衡处理原始训练集上)和SCRF方法(即对原始训练集随机增加少数类的样本从而得到平衡的训练集,再将CRF模型应用在平衡后的训练集上)的表现,通过这个实验可以知道混合高斯分布伪样本生成技术对情感分析任务的贡献.(2) 比较GWCRF和GCRF方法的表现,通过这个实验可以看出利用Word2vec扩展微博对于情感分析任务的贡献.(3) 比较GWCRF、SVM和BP神经网络方法的表现,通过这个实验可以看出GWCRF方法对于情感分析任务的有效性.

对每个数据集,使用10折交叉验证评估算法的性能.为了更直接和更公平地比较本文方法和其他方法的结果表现,本文使用少数类召回率PY(即预测正确的少数类样本占少数类总样本的比例)、多数类召回率PN(即预测正确的多数类样本占多数类总样本的比例)和G-mean = $\sqrt{PY \times PN}$ 作为本文的评价指标.一个好的方法应该在两个类别(这里是指少数类和多数类)上都有很好的效果,从而会有更高的G-mean.

3.2 实验数据集

本文从新浪微博API获取大约3万条未标记语料作为训练GWCRF模型的数据.过滤掉一些广告文本后,得到表2的数据集,这些微博包含了四个热门主题,都是带有明显情感偏向(负面情感比正面情感多)的社会事件,例如(#90后暴打老人#、#食用油涨价#等),这些主题的情感倾向都是不平衡的.本文使用这四个不平衡数据集作为实验数据,分别用 D_1 到 D_4 来表示.从 D_1 到 D_4,数据集的情感倾向分布不平衡性依次递增.

表2 数据集的详细情况

Tab.2 The details of data set

数据集	正面情感句子数量	负面情感句子数量	正面情感句子数量/句子总数
D_1	338	1 216	0.278
D_2	135	983	0.137
D_3	72	1 302	0.055
D_4	69	2 531	0.027

3.3 实验结果与分析

由表3和表4可知GCRF方法在四个数据集的实验结果都优于CRF方法,在少数类召回率上平均提升6.8%,在G-mean上平均提升5%,这说明利用混合高斯分布伪样本生成技术平衡训练集后能使情感识别效果有明显的提升.此外,CRF方法对少数类样本的情感倾向预测性能较低,而它对于多数类样本预测性能较高,这是数据集情感倾向分布不平衡导致的.传统方法对正面情感样本和负面情感样本同等处理,而未考虑在实际数据集中情感倾向分布不平衡的影响.当训练数据集中情感倾向分布相差悬殊时,预测结果明显偏向多数类样本,从而导致少数类样本的预测精度降低.

另外,从表4和表5中可以看到GCRF方法的

结果比 SCRF 方法在 G-mean 指标上平均提升 3.3%,这说明利用混合高斯分布为少数类增加伪样本是有效的,它能够增加很多有利于分类器判别的信息.

表 3　CRF 方法的实验结果

Tab. 3　The experimental results of CRF approach

PY	PN	G-mean	
D_1	0.101	0.986	0.315
D_2	0.076	0.981	0.273
D_3	0.000	1.000	0.000
D_4	0.000	1.000	0.000

表 4　GCRF 方法的实验结果

Tab. 4　The experimental results of GCRF approach

PY	PN	G-mean	
D_1	0.714	0.862	0.784
D_2	0.723	0.806	0.764
D_3	0.705	0.798	0.751
D_4	0.622	0.761	0.688

表 5　SCRF 方法的实验结果

Tab. 5　The experimental results of SCRF approach

PY	PN	G-mean	
D_1	0.415	0.521	0.465
D_2	0.430	0.562	0.491
D_3	0.357	0.339	0.349
D_4	0.320	0.312	0.315

SCRF 方法效果不好的原因是随机增加少数类样本的方式很可能出现样本重叠的情况,从而不能很好模拟数据集的分布. 在表 4 和表 6 中,可以看到在 4 个数据集上 GWCRF 的结果平均比 GCRF 提升了 1.1%,这证明了利用 Word2vec 来扩展微博能够丰富微博句子的情感信息,从而有利于提高情感分类任务的性能. 例如,句子“它的屏幕很细腻”,假如情感词典中没有“细腻”这个情感词,这时就识别不出这个句子是正面的. 然而,假如利用 Word2vec 求得它的相似词是“精致”,并且情感词典中有该词,这时就可以识别出句子的情感.

从表 7 中可以看出 GWCRF 方法在中文微博情感分析任务上比 SVM 方法和 BP 神经网络方法取得更好的效果,说明 CRF 模型应用在经过本文混合高斯分布伪样本生成技术和 Word2vec 技术处理后的数据上时能提高预测性能. 因为 CRF 模型不仅能够处理复杂的特征,而且能够引入句子间的上下文依赖信息.

表 6　GWCRF 方法的实验结果

Tab. 6　The experimental results of GWCRF approach

PY	PN	G-mean	
D_1	0.874	0.897	0.885
D_2	0.820	0.913	0.867
D_3	0.812	0.872	0.841
D_4	0.803	0.855	0.830

表 7　SVM、BP 神经网络、GWCRF 方法的实验结果

Tab. 7　The experimental results of SVM, BP and GWCRF approach

data	G-mean		
	SVM	BP	GWCRF
D_1	0.485	0.451	0.885
D_2	0.443	0.433	0.867
D_3	0.416	0.419	0.841
D_4	0.308	0.311	0.830

从表 3 到表 7 可知,从 D_1 到 D_4,随着数据集情感倾向分布不平衡性的增加,各个方法的召回率和 G-mean 值都会出现不同程度的下降,而本文提出的 GWCRF 方法依然能取得不错的效果,这证明了 GWCRF 方法在情感倾向分布不平衡的中文微博数据集的情感倾向识别问题上是有效的.

4　结束语

本文提出了一种处理情感倾向分布不平衡的中文微博数据集的情感倾向识别问题的方法. 该方法包含了混合高斯分布伪样本生成技术和 CRF 预测模型. 混合高斯分布伪样本生成技术中,通过增加伪样本的方式增加少数类样本的数量来平衡训练集. 在预测模型中,首先利用 Word2vec 来扩展微博,然后将 CRF 模型应用在平衡和扩展后的训练集上. 实验结果证明, GWCRF 方法在中文微博情感分析问题上能比传统的方法取得更好的效果.

参考文献:

[1] XIE L, ZHOU M, SUN M. Hierarchical structure based hybrid approach to sentiment analysis of chinese micro blog and its feature extraction[J]. Journal of Chinese Information Processing, 2012, 26(1): 73-83.

[2] VHUTTO C J, GILBERT E. VADER: A parsimonious rule-based model for sentiment analysis of social media text[C]//Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media. Phoenix, Arizona, USA: Association for the Advancement of Artificial Intelligence, 2014: 216-225.

[3] PANDARACHALIL R, SENDHILKUMAR S, MAHALAKSHMI G S. Twitter sentiment analysis for large-scale data: an unsupervised approach [J]. Cognitive Computation, 2014, 7(2): 254-262.

[4] ZHOU S, CHEN Q, WANG X. Active deep learning method for semi-supervised sentiment classification [J]. Neurocomputing, 2013, 120(10): 536-546.

[5] 吴江,唐常杰,李太勇,等. 基于语义规则的 Web 金融文本情感分析[J]. 计算机应用,2014,34(2):481-485.
WU J, TANG C J, LI T Y, et al. Sentiment analysis on Web financial text based on semantic rules[J]. Journal of Computer Applications, 2014, 34(2):481-485.

[6] 李寿山,李逸薇,黄居仁,等. 基于双语信息和标签传播算法的中文情感词典构建方法[J]. 中文信息学报,2013,27(6):75-81.
LI S S,LI Y W,HUANG J R, et al. Construction of Chinese sentiment lexicon using bilingual information and label propagation algorithm [J]. Journal of Chinese Information Processing, 2013, 27(6): 75-81.

[7] TANG D, WEI F. Building large-scale Twitter-Specific sentiment lexicon: a representation learning approach[C] // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. Dublin, Ireland: Technical Papers, 2014: 172-182.

[8] BARBOSA, FENG J. Robust sentiment detection on Twitter from biased and noisy data. [C] // Proceedings of COLING 2010, the 23rd International Conference on Computational Linguistics. Beijin, China: Posters Volume, 2010: 36-44.

[9] VANZO A, CROCE D, BASILI R. A context-based model for sentiment analysis in Twitter[C] // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. Dublin, Ireland: Technical Papers, 2014: 2345-2354.

[10] 杨经,林世平. 基于 SVM 的文本词句情感分析[J]. 计算机应用与软件,2011,28(9):225-228.
YANG J, LIN S P. Emotion analysis on text words and sentences based on SVM[J]. Computer Applications and Software, 2011, 28(9) : 225-228.

[11] 陈培文,傅秀芬. 采用 SVM 方法的文本情感极性分类研究[J]. 广东工业大学学报,2014(3):95-101.
CHEN P W, FU X F. Research on sentiment classification of texts based on SVM[J]. Journal of Guangdong University of Technology, 2014(3): 95-101.

[12] 滕少华,吴昊,李日贵,等. 可调多趟聚类挖掘在电信数据分析中的应用[J]. 广东工业大学学报,2014(3):1-7.
TENG S H, WU H, LI R G , et al. The application of the adjustable multi-times clustering algorithm in telecom data sentiment analysis on web financial text based on semantic rules[J]. Journal of Computer Applications, 2014(3): 1-7.

[13] DAVIDOV D, TSUR O, RAPPOPORT A. Enhanced sentiment learning using Twitter hashtags and smileys[C] // Proceedings of COLING 2010, the 23rd International Conference on Computational Linguistics. Beijing, China: Posters Volume, 2010: 241-249.

[14] JIANG L, YU M, ZHOU M, et al. Target-dependent Twitter sentiment classification[C] // The Meeting of the Association for Computational Linguistics, Human Language Technologies, Proceedings of the Conference. Portland, Oregon, USA: Association for Computational Linguistics, 2011: 151-160.

[15] TASKAR B, ABBEEL P, KOLLER D. Discriminative probabilistic models for relational data[J]. Eprint Arxiv, 2012, 7(3): 485-492.