一种基于半监督学习方法的问答对广告判定系统及演示

陈修司 李笑宇 高 军

（北京大学信息科学技术学院 北京 100871）

（高可信软件技术教育部重点实验室（北京大学） 北京 100871）

（xiusichen@pku.edu.cn）

Demo of a Semi-Supervised Method Based System to Detect Advertisements from QA pairs

Chen Xiusi, Li Xiaoyu, and Gao Jun

(School of EECS, Peking University, Beijing 100871)

(Key Laboratory of High Confidence Software Technologies(PKU), Ministry of Education, Beijing 100871)

Abstract With the development of the internet, the QA online communities take more important roles in acquiring knowledge and solving problems. The users in these communities utilize the services to perform operations like asking questions so as to acquire satisfying answers. However, some advertisers tend to post advertisements for some commercial purposes in the form of answer content, which lowers the effectiveness of the whole system. We developed algorithms to discriminate advertisements from normal answers, in which conventional machine learning methods like Logistic Regression, Random Forest, Support Vector Machine, and ensemble methods involving these algorithms are in our concern.

Key words Advertisement Detection; QA community; Semi-supervised; Logistic Regression; Random Forest

摘要 互联网互动问答社区已成为网民获取知识、解决问题的重要手段。但是，一些投机用户利用问答形式，塞入广告或相关的推广内容，试图欺骗搜索引擎。因此，发现并剔除这种广告数据，是保证社区健康发展的重要环节。本文的目标就是根据问答系统所提供的真实广告数据，设计和实现问答系统中的广告推广数据的挖掘算法，为这一环节提供技术支持。我们通过建立一个只有正例和无标注数据的分类器来达到这一目标。这个分类器利用了传统的机器学习算法：逻辑斯蒂回归、随机森林、支持向量机，等等，以及它们的组合，来构造最终的分类器，并取得了良好的效果。

关键词 广告检测；问答系统；半监督学习；逻辑斯蒂回归；随机森林

中图法分类号 TP274

随着互联网的发展，互动问答社区已成为网民获取知识、解决问题的重要手段和知识沉淀的重要场景，影响力日益提高。在互动社区中用户利用互联网服务商提供的问答系统进行提问、回答、采纳等操作，获取对于问题的满意答案。搜狗问问[1]、百度知道[2]是国内最为大家熟知的问答系统，这两家问答系统都已经产生数亿的“问题答对”，这些数据可用于帮助广大网民解决生活中各种各样的问题。

[1] 搜狗问问：http://wenwen.qq.com/

[2] 百度知道：https://zhidao.baidu.com/

收稿日期：2016-06-30

基金项目：国家自然科学基金项目（61272156、61572040）；深圳市科技计划项目 JCYJ20151014093505032

但是，问答系统面临着一个必须解决的难题:一些投机的用户出于商业目的，利用提问与回答的形式，塞入广告或相关的推广内容，试图欺骗搜索引擎，获取私利，这就干扰了社区秩序，破坏了用户体验。因此，发现和剔除这种为了商业目的而编写的欺诈数据，就成为净化系统生态环境，保证社区健康发展的重要环节。为此，搜狗公司在第三届中国数据库学术大会万维网知识提取竞赛中，确定“问答系统的广告推广数据挖掘”为主题。本文的目标就是根据问答系统所提供的真实数据，设计和实现问答系统中的广告推广数据的挖掘算法，为这一环节提供技术支持。

目前针对广告识别问题已经展开研究工作。其中 Chen C, Wu K, Srinivasan V, et al.的工作[1]将最佳答案抽取出来，只考虑问题与最佳答案构成的问答对是否为广告。他们提出三个指标，分别衡量了特定用户分别作为提问者与回答者与广告相关的可能性、一个特定问答对的文本与广告相关的可能性。对于每个问答对获取以上 3 个指标后，采用 Logistic Regression 训练得到一个分类器。

Li X, Liu Y, Zhang M, et al.则尝试从不同的角度入手[2]，认为任何推销者想要在回答中推销自己的产品或是服务，都必须在回答中留下“推销途径”（promotion channel），“推销途径”包括 URL，电话号码或者其他社交网络账号。作者想要从推销途径入手来找出有广告推销嫌疑的回答，利用用户的社区属性，将待判断的问答对构造“推销途径-回答者”二部图，选取最有可能是推销途径的“途径点”作为种子点，将种子点的权重通过传播算法传达整个二部图。以上过程结束后，获得提问者和回答者在二部图中对应结点的权重，以及该问答对回答内容中出现的所有推销途径在二部图中权重最大的点对应的权重。最后利用 Logistic Regression 得到分类器。

以上工作都已经得到了正例和负例训练集。然而在本次知识提取竞赛中，我们只能得到正例集合，即确定为广告的训练数据。对于问答社区的审查人员，他们更加关注那些确定是广告的记录。当审查到一条广告推广信息时，会立刻将其标注为广告。相反，由于广告信息往往各式各样，隐藏较深，所以并不能直接确定某一问答对不含有广告，以致于直接标注某一条不确定问答对为非广告往往具有比较低的置信度。也就是说，我们对于所发现的正例情况很有信心，而对反例结果不是很有信心。这时候把这些数据看成是无标注数据比看成反例数据更为合适。此时，大多数已有解决方案无法直接套用到我们面对的这种情况。

另外，由于广告形式各异，直接可用特征数量非常有限，我们需要从可以获取的各种形式的数据中深入分析挖掘各广告相关的特征，并试验其有效性。

本文的工作对于给定的只有正例的真实广告数据，进行问题分析和建模，并提出了一种新的方法，并在固定目标为 F-value 的情况下利用实验验证了本文提出的方法在真实广告数据上能够获得优良的效果。

接下来的第 1 部分将介绍广告数据特征的提取；第 2 部分将描述分类方法；第 3 部分，给出实验结果；第 4 部分将是系统演示的简述；最后将对本文加以总结。

1 广告数据特征提取

如上文所述，提取与广告高度相关的特征对于训练出效果优良的分类器具有至关重要的作用。本节重点介绍我们对于任一问答对提取的各个特征。

1.1 文本特征

问答对本质上由问题文本与回答文本构成，推广者要实现推广目的，无可避免地要利用问答文本，故提取文本特征是最直观的想法。由于我们处理的数据为中文问答社区真实数据。为了正确提取文本特征，我们首先对问答文本进行了中文分词。

得到经过分词的文本后，常见的提取文本特征的方式包括：1.提取 TF-IDF 特征；2.利用深度学习提取词嵌入特征。二者思想的侧重点不尽相同。TF-IDF 特征侧重考虑文档集中的词汇的区分度，将每个词对于每篇文档的独特性纳入计算特征值的考虑范围。词嵌入则着重考虑词与词之间的相对语义关系，它使得相似的文本拥有相似的词嵌入向量。

我们的方法同时利用了以上两种文本特征。

tfidf	在确定为广告的数据集中提取词频最高的 300 个词，只计算这些词的 TF-IDF 特征
word2vec	word embedding 特征，将每一个在训练数据中出现过的词映射到 300 维特征空间中

表格 1 文本特征

注意到我们利用两种方式都是提取至 300 维向量。因为在我们的实验过程中发现 TF-IDF 特征维数过高可能会引起过拟合；而利用 word embedding 将词映射到过高维度的向量则得到的 Top-K 近邻结果与实际结果不准确。

1.2 用户特征

社区产品最大特点是有人的属性以及行为。尤其在问答系统中，用户之间的关联性非常强。在识

别广告推广数据时，不仅文本特征可用，用户的行为以及用户之间关系都可以用于广告推广数据的挖掘。

在问答社区中，我们可以获得的用户属性包括问答者的：ID、昵称、IP 地址等信息。经过分析挖掘，我们从这些信息中提取的特征包括：

is_promotion_account	提问者和回答者昵称中是否带有“网”或“代理”关键字
is_self	提问者与回答者是否是同一个账号
is_same_LAN	提问者与回答者 IP 是否在同一个网段

表格 2 用户特征

1.3 问答对相关特征

一个问题与对于它的某一个回答，共同构成一个问答对。对于任一问答对，我们可以得到的信息包括：1.问题 ID 与回答 ID；2.提问者悬赏分数；3.提问时间与回答时间；4.提问者、回答者是否匿名；5.提问者是否满意该条回答。

通过分析搜狗提供的测试数据，对于这些信息，我们提取或直接利用的特征包括：

Δt	提问时间 – 回答时间 是否小于 δ
anonymous_or_not	提问者是否匿名、回答者是否匿名

表格 3 问答对特征

1.4 推广渠道相关特征

受到[2]的启发，我们认为推广途径确实是广告的一个必要条件，但是我们并没有直接构建用户-推广途径二部图，而是将各种推广途径并入我们的特征：

在实际处理数据时，我们遇到并解决了一些棘手的问题，如：中英文组合的联系方式、利用谐音字或形似字代替联系方式中常见关键字，以及只有具有某一些特征的 URL 才能被判定为广告等。

pattern_numberseries	回答中是否含有手机号、QQ、微信等
pattern_URL	回答中是否含有 URL

表格 4 推广渠道相关特征

2 广告分类方法

如前文所述，我们处理的是一个 PU(Positive Unlabeled)学习问题。我们采用三种分类方法来建立 PU 分类器：两步方法、直接方法以及组合方法。

2.1 直接方法

Liu B, Dai Y, Li X, et al.介绍了一种适用于文本分类的经过修改的 SVM 方法[3]，称为偏置 SVM（Biased-SVM）。这个方法通过稍微修改 SVM 的公式使得它适用于 PU 学习。具体来说，假设前k – 1个数据实例是正例集合P，其余部分是无标注集合U。先把U中数据全部标注成反例，加入P共同构成训练集合。通过在 SVM 优化目标函数中对正例P错误与无标注数据U错误赋予不同的惩罚权重来训练出一个好的分类器：

假设分割平面为：

$$f(\mathbf{x}) = < \boldsymbol{\omega} \cdot \boldsymbol{x} > + b = 0$$

此时我们需要最小化：

$$\frac{< \boldsymbol{\omega} \cdot \boldsymbol{\omega} >}{2} + c_+ \sum_{\mathrm{i}=1}^{k-1} \xi_i + c_- \sum_{i=\mathrm{k}}^{n} \xi_i$$

满足条件：

$$y_i(< \boldsymbol{\omega} \cdot \boldsymbol{x_i} > + b) \geq 1 - \xi_i, i = 1, 2, \ldots, n$$

$$\xi_i \geq 0, i = 1, 2, \ldots, n$$

在实际操作中，我们给c_+赋予一个较大值，c_-较小值，因为在被看成是反例集合的无标注集合中含有正例数据，对于这样的“脏”数据容忍度要相对高一些。

2.2 两步方法

顾名思义，该方法由两个步骤组成：

（1）从无标注数据集U中发现一些可靠的反例文档集合（Reliable Negative），用RN表示。在这一步中，我们利用 2.1 中介绍的偏置 SVM 的方法。通过调整C_+和C_-，来使得选出的RN尽量不含有广告数据。

（2）利用正例P、RN、U－RN来建立分类器。根据RN集合中数据的质量和数量的不同，在这个步骤中可能会使用某个学习算法一次或循环多次。在这一步中，我们尝试了包括 SVM，Logistic Regression，Random Forest 等分类方法。

不难发现，两步方法的思路就是将 PU 学习转化为经典的机器学习问题来处理，关键步骤就是筛选出一部分非常可靠的负例数据。

两步方法能带来的额外好处是，我们通过第（1）步获得的同时具有正例和负例的训练数据可以用来进行交叉验证，使得我们可以搜索在验证集合上效果最好的模型参数。

2.3 组合方法

Dietterich 的工作[4]通过三个角度阐述了组合方法的有效性。

在我们的工作中有两处用到了组合方法思想。在两步方法的第（2）步中，Random Forest 方法在构造每一棵决策树时随机抽取若干特征，一条测试数据的最后分类由森林中的所有决策树共同决定。

同时根据[4]，由于不同的分类模对于分类结果的误差分布的假设不同，故将它们的结果综合考虑将消除一部分单一模型带来的随机误差。在我们以尽可能高的 F-value 为目标的前提下，组合方法能够在保证精确度不降低的情况下提高召回率。故将

Logistic Regression，SVM，Random Forest 都考虑进我们的模型是一个好的想法。最终，我们的分类模型结构如图 1。

对于任意待预测数据X_i，分别用训练好的 SVM、Logistic Regression、Random Forest 进行分类，对每种模型赋予一定的权重，若三种模型权重之和大于η（一般取权重之和的一半），则判断为正例（广告），否则为负例（非广告）。

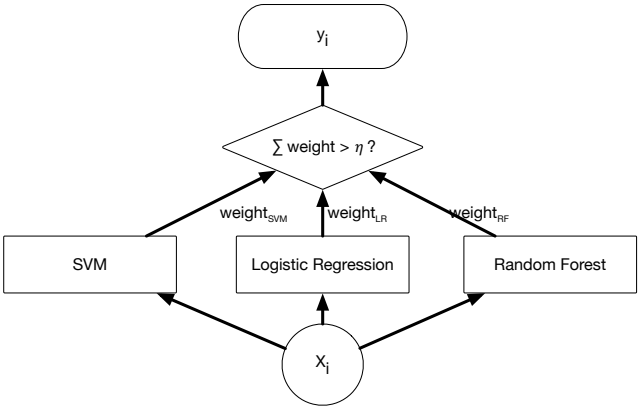

图 1 模型层次结构

3 测试效果

大赛官方给定的优化目标为一个加权的 F 值：

$$F = \frac{1}{\frac{1.5}{precision} + \frac{1}{recall}}$$

以上特征及分类方法是根据参加竞赛过程中多次提交反馈进行优化的结果。待分类广告总量为 1050954 条。最终提交并获得第二名的结果：

精确度	召回率	F 值
0.6086	0.4559	0.5616

表格 5 最终测试结果

4 系统架构和系统演示

本节描述系统体系架构、演示步骤及环境。

4.1 系统架构

本文提出的广告判定系统架构如下图：

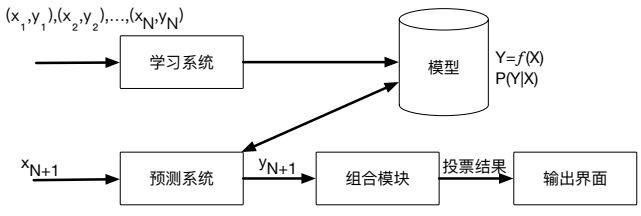

图 2 系统架构

学习系统利用输入数据训练出多个分类模型，新来的待分类数据作为每个分类模型的输入，分类模型输出结果由组合模块进行组合，最后将判定结果传到输出界面进行展示。

4.2 系统演示

4.2.1 系统演示步骤

如图 4，我们将首先对我们处理的数据进行展示，在将原始数据提取特征后，我们将把中间结果文件进行展示。这样做的好处是，可以看到诸如 TF-IDF 特征的每一维所代表的词这样的对应关系，也可以重现我们验证各特征有效性的过程；其次，利用预先训练好的模型进行分类以及各模型组合投票的过程也将进行可视化展示；原始待分类数据和它们对应的分类结果将在最后进行展示。

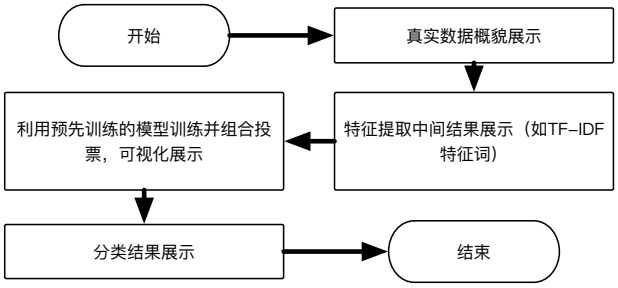

图 3 演示步骤

4.2.2 系统演示环境

操作系统：Mac OS X 或 Linux

Python：2.7 以上或 3.5 以上

Python 依赖库：Scikit-Learn, Gensim, Matplotlib

5 结束语

本文将问答社区的广告判定问题建模为一种只有正例和无标注数据的半监督学习问题，并针对这个问题提出了一种新的解法：利用偏置 SVM 方法从无标注数据集中筛选出可靠的负例数据，并与已有的正例数据合并组成第二轮学习的训练数据。第二轮训练阶段，利用 SVM、Logistic Regression、Random Forest 模型及其组合方法，使得精确度不损失、召回率有所提高。最后，给出了整个分类（判定）系统的体系结构和演示流程。

参 考 文 献

[1] Chen C, Wu K, Srinivasan V, et al. The best answers? think twice: online detection of commercial campaigns in the CQA forums[C]//Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013: 458-465.

[2] Li X, Liu Y, Zhang M, et al. Detecting promotion campaigns in community question answering[C]//Proceedings of the 24th International Conference on Artificial Intelligence. AAAI Press, 2015: 2348-2354.

[3] Liu B, Dai Y, Li X, et al. Building text classifiers using positive and unlabeled examples[C]//Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003: 179-186.

[4] Dietterich, T. G.[R] Ensemble Methods in Machine Learning. In J. Kittler and F. Roli (Ed.) First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science (pp. 1-15). New York: Springer Verlag, 2000

陈修司，男，1993 年生，硕士研究生，主要研究方向为数据库系统、数据挖掘算法。

李笑宇，男，1995 年生，本科生，主要研究方向为数据挖掘算法。

高军，男，1975 年生，教授，博士生导师，主要研究方向为数据库系统、数据挖掘算法、推荐系统。

NDBC 2016 审稿意见：

无审稿意见，直接录用

作者联系方式：

陈修司

手机：+86 188 1052 1442

邮箱：xiusichen@pku.edu.cn

地址：北京市海淀区颐和园路 5 号北京大学理科一号楼 1636

邮编：100871