软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(Suppl.(2)):127−135

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

电子商务网站中误导性商品描述识别*

龙 吟[1,2], 刘红岩[3], 何 军[1,2], 胡 鹤[1,2], 杜小勇[1,2]

[1](数据工程与知识工程教育部重点实验室(中国人民大学),北京 100872)

[2](中国人民大学 信息学院,北京 100872)

[3](清华大学 经济管理学院,北京 100084)

通讯作者: 何军, E-mail: hejun@ruc.edu.cn

摘 要: 网上购物已被越来越多的消费者接受,C2C 网站作为主流购物平台提供数以万计的商品条目供消费者选择,其中有一定数量商品条目的商品描述具有误导性.误导性是指条目的商品描述与其实际价格不符合,通常的表现是描述商品的价格低于其应有的价格,以此吸引消费者,误导消费者到其购物页面.这既影响消费者的判断,又损坏购物网站的信誉度.为了找出这部分具有误导性的商品描述,提出了一种结合概率模型 HMM 和基于统计的异常值识别方法,能够有效地识别出误导性商品描述.HMM 模型从概率的角度有效地确定商品描述所指代的商品,为 C2C 网站上商品描述的不规范导致的商品指代信息模糊提供了一种行之有效的解决方法.基于统计的异常值识别方法在处理 C2C 网站上商品信息比较单一时较为有效.用该方法在实际的电商网站数据集上进行了实验.实验结果证明了该方法的有效性.

关键词: 误导性描述;HMM;异常值检测

中文引用格式: 龙吟,刘红岩,何军,胡鹤,杜小勇.电子商务网站中误导性商品描述识别.软件学报,2014,25(Suppl.(2)):127−135. http://www.jos.org.cn/1000-9825/14031.htm

英文引用格式: Long Y, Liu HY, He J, Hu H, Du XY. Identification of misleading product description in e-commerce website. Ruan Jian Xue Bao/Journal of Software, 2014,25(Suppl.(2)):127−135 (in Chinese). http://www.jos.org.cn/1000-9825/14031.htm

Identification of Misleading Product Description in E-Commerce Website

LONG Yin[1,2], LIU Hong-Yan[3], HE Jun[1,2], HU He[1,2], DU Xiao-Yong[1,2]

[1](Key Laboratory of Data Engineering and Knowledge Engineering of the Ministry of Education (Renmin University of China), Beijing 100872, China)

[2](School of Information, Renmin University of China, Beijing 100872, China)

[3](School of Economics and Management, Tsinghua University, Beijing 100084, China)

Corresponding author: HE Jun, E-mail: hejun@ruc.edu.cn

Abstract: Online shopping has been accepted by more and more consumers. C2C websites provide thousands of offers for consumers as a mainstream e-commerce platform. When customers search products in C2C website, some returned offers have misleading description. Misleading description means that the description does not convey the actual price of products, but usually claiming much lower price for the purpose of attracting more consumers. The misleading offers affect consumers' judgments and bring bad influences on the websites' reputation. This paper proposes an approach that combines statistical model HMM with statistical outlier detection method to detect misleading offers. HMM model is built to determine the product that an offer description really designates, providing an efficient solution to eliminate the ambiguity of the offer description caused by description irregularities. The statistical outlier detection method is effective to deal with limited product offer information. The paper further conducts experiments on real data set of electric business websites and the results demonstrate the effectiveness of the proposed approach.

* 基金项目: 国家自然科学基金(71272029, 71110107027); 国家社会科学基金(12&ZD220); 国家高技术研究发展计划(863)(2014AA015204)

收稿时间: 2014-05-07; 定稿时间: 2014-08-19

Key words: misleading description; HMM; outlier detection

1 研究背景及动机

网络的出现在一定程度上影响着人们的生活、工作和学习,而基于互联网发展起来的电子商务引领了消费的新潮流,近年来伴随着网络科技的发展,各类电商网站发展迅速,电商网站的模式主要分为B2C,C2C等类型[1].

B2C是指商家直接面向消费者销售产品和服务.中国国内采用B2C模式的网站主要有新蛋网、京东、亚马逊等.B2C网站货源可靠,行销成本低,既保证了商品质量,又压低了商品价格.所以,商品在B2C网站上的价格可作为该商品的标准价格.C2C是指个人与个人之间的电子商务.中国国内采用C2C模式的网站主要有淘宝网、易趣网、拍拍网等.由于C2C网站是开放的贸易平台,任何人都可以注册成为卖家,这样的低门槛使得大量的卖家提供同样的商品,因此出现同质化竞争激烈的情况,同样商品的价格也有高有低,分布范围较广,有部分商家恶意竞争,使网站上出现了一定数量的误导性商品条目.

商品条目是指卖家在网站上对商品进行的概要展示,条目包括商品图片、商品文字描述、商品价格以及购买该商品的页面连接.淘宝网的搜索结果是以商品条目列表的形式显示不同条目的信息,如搜索“佳能 650D 18-55”得到的结果如图1所示.在图1的第2个商品条目中,描述是“佳能单反EOS 650D相机 650D/18-55 套机原装正品实体经营”,价格是“2430 元”,但在该条目的购买页面里,下订单处有诸多选项,价格“2430 元”所对应的商品是佳能“650D 相机单机”,并不是描述表达的相机套机.

Fig.1 List of product items at C2C Website

图1 C2C网站商品条目列表

除了上述情况即“条目价格是较低价配置的价格,描述的是较高价格的配置”外,还有多种情况.表1列出了误导性商品描述的类型,如果某条目出现表1所述的3种情况之一,则该条目的描述是具有误导性的.本文将该条目的描述称为误导性描述,将该条目称为误导性条目.商家用误导性条目误导用户,增加其商品页面的曝光率,从而提高交易量,既劣化了消费者的用户体验,也降低了网站的信誉程度.因此,识别出误导性商品描述非常有必要.

Table 1 Categories of misleading product description

表1 误导性商品描述的类型

	描述	价格
情况1	较高配置的商品	较低配置的价格
情况2	行货商品	水货商品的价格
情况3	全新商品	折旧商品的价格

2 相关工作

关于C2C网站的问题研究由来已久,但大多是对欺诈问题的研究.郑华等人提出了一种能够识别电子商务中信誉欺诈的方法.该方法用神经网络和SNA网络结合起来判断欺诈行为[2].朱艳春等人对C2C网站回馈评分

中的欺诈行为进行了研究,同样从社交网络的角度去研究[3].与以上研究不同,我们的研究主要关注的是误导性商品描述.这个问题还很少有人研究.

异常值是影响数据质量的一个非常重要的因素.一直以来,科学研究者提出各类方法来解决异常值识别的问题.基于密度的异常值识别方法是常用方法之一,由Breunig等人提出的考虑数据局部密度的异常值识别算法LOF[4]是近年来被大量学者认可的一种方法.LOF提出局部异常值因子(local outlier factor),算法计算每个对象的异常值得分,认为分数越高的对象越有可能是异常值,但确定这些异常值得分的上下界以判断对象是否是异常值是一个较难的问题.基于密度的观点来说,异常点是在低密度区域中的对象.在本文所要解决的问题中,不同商品的价格虽然分布有所差异,但每种商品的误导性条目与其余条目的价格分布都较为均匀,密度差别不明显,因此,基于密度的异常值识别方法不适用于本文所要解决的问题.基于聚类的异常值识别方法也是较为常用的方法,DBScan[5]就是其中之一.基于聚类的观点来说,如果某对象不属于任何类,则该对象是异常点.在本文所要处理的数据中,误导性条目的价格数值上较低,但与其余商品条目相比,并无明显特点,对于聚类方法,可用Feature较为单一,因此,这类方法同样不适用于本文所要解决的问题.本文根据所要处理的数据的特点,使用了一种基于统计的异常值识别方法,能够较好地解决误导性描述识别的问题.

3 误导性商品描述的识别方法

3.1 问题定义

当在C2C网站上搜索一种商品时,会返回很多商品条目,对应商家提供的不同商品描述,这些条目中包含的价格有高有低,绝大部分条目的商品描述与其实际售价是符合的,但有小部分价格较低的商品条目是误导性条目.本文要解决的问题是:给定一个商品的品牌和型号,在C2C网站上搜索返回的商品描述中识别出所有有误导性的描述.

3.2 解决方案的基本框架

本文提出的解决方案主要由3个步骤构成:步骤1,将所有商品条目的商品描述分解、识别,确定条目的商品描述究竟指的是什么商品;步骤2,将商品描述指的同一商品的条目分到一个商品组;步骤3,若商品组里有误导性条目,则找出误导性条目.算法1对以上步骤做了总结.

算法1. 误导性商品描述识别.

输入:C2C网站上的商品条目.

输出:误导性商品描述.

主要步骤:

1. 对每条商品条目,进行步骤2.
2. 将一个商品条目的商品描述中的关键字段抽取出来,若有多种抽取结果,则用隐马尔可夫模型HMM确定最优的抽取结果.从抽取出的关键字段得到商品描述所指的商品.
3. 将所有映射到同一商品的商品条目分到一个商品组.
4. 若商品组内有误导性商品描述,则通过统计方法找出误导性商品描述.

第3.4节、第3.5节介绍了算法中每个步骤的详细过程.

3.3 隐马尔可夫模型HMM

HMM[6]是一个双重随机过程,即HMM由两个随机过程组成:一个是隐含的状态转移序列,它对应一个单纯的Markov过程;另一个是状态决定观测的随机过程.并且,在这两个随机过程中,有一个随机过程(状态转移序列)是不可观测的,只能通过另一个随机过程的输出观测序列进行推断,所以称为隐马尔可夫模型.隐马尔科夫模型有3个假设:有限历史性假设——系统在t时刻的状态只依赖于$t-1$时刻的状态;输出独立性假设——t时刻所生成的观测值只依赖于t时刻的状态;齐次性假设——状态与具体的时间无关.

设HMM的状态集为$S=\{s_1,s_2,...,s_N\}$,观测集为$V=\{v_1,v_2,...,v_M\}$,模型在时间t的状态记为$q_t,q_t\in S,1\leqslant t\leqslant T,T$

是观察序列的长度.模型记录的状态序列记为 $Q=\{q_1,q_2,...,q_t\}$.一个隐马尔可夫模型的基本参数包括:

(1) 状态转移的概率分布 A

状态转移的概率分布可表示为 $A=\{a_{ij}\}$,其中,$a_{ij}=P\{q_{t+1}=s_j|q_t=s_i\}$,$1\leqslant i,j\leqslant N$,且满足 $a_{ij}\geqslant 0$, $\sum_{j=1}^{N}a_{jj}=1$, 表示时刻 t 从状态 s_i 转移到时刻 t+1 状态 s_j 的转移概率.

(2) 状态 s_i 条件下输出的观测变量概率分布 B

假设观测变量的样本空间为 V,在状态 s_i 时输出观测变量的概率分布可表示为 $B=\{b_i(v),1\leqslant i\leqslant N,v\in V\}$,其中,$b_i(v)=f\{Q_t=v|q_t=s_i\}$,$Q_t$ 为时刻 t 的观测随机变量.可以是一个数值或向量,观测序列记为 $O=\{O_1,O_2,\ldots,O_t\}$.值得注意的是,此处观测变量的样本空间和概率分布可以是离散型,也可以是连续型.

(3) 系统初始状态概率分布π

系统初始状态分布可表示为$\pi=\{\pi_i,1\leqslant i\leqslant N\}$,其中$\pi_i=P\{q_i=s_i\}$.

综上所述,要描述一个完整的 HMM,除基本模型参数、状态数 N 和观测值数 M 以外,还需要参数 A,B,π,即隐马尔科夫模型可形式化定义为一个五元组$\lambda=\{N,M,A,B,\pi\}$.对于一个标准的 HMM 模型,可解决 3 类问题:

- 评估问题——给定λ,根据前向后向算法(forward-backward)来求某一输出观测值序列 O 的概率 $P(O|\lambda)$;
- 解码问题——给定λ和观测字符序列 O,根据 Viterbi 搜索算法可以找到产生这一序列概率最大的状态序列;
- 学习问题——给定λ和 O,调整模型参数,使得产生这一序列的概率最大.

3.4 商品描述的分解和识别

商品条目的商品描述中文字描述格式不统一,含有多种信息,描述中可能包含不止一个商品的品牌和型号.因此首先需要将文字描述分解,抽取商品的品牌和型号,识别该条目的商品描述指的是什么商品.我们从 B2C 爬取商品的型号,品牌以及价格,用爬取的数据建立一个商品名称的列表 Product List,内容包括商品的“品牌”和“型号”,其中每条记录以品牌在前型号在后的有序形式记录,将这种形式命名为“B+M”.每个商品有唯一的一条记录,其中型号部分可能由多个字段组成,例如“佳能 700D 18-105”,型号部分由“700D”和“18-105”组成.与此同时,建立了一个商品标准价格的列表 Product Price List,该列表的记录与 Product List 中的记录一一对应.用 Product List 中的记录作为搜索词到 C2C 网站上获取商品的搜索结果,即商品条目.商品条目的描述中一定包含搜索词,但也可能提及 Product List 中没有的商品信息.我们将 Product List 中没有,而搜索结果的商品描述中包含的商品称为新商品.根据搜索结果中的描述,我们归纳了一个 Brand List,Brand List 包含 Product List 中出现的品牌以及在商品描述中出现的新商品的品牌.同时,根据 Product List 中已有的商品和新品牌的商品信息,我们归纳了所有可能出现在商品条目中的商品型号的命名组成方式(见表 2)和商品名规则(见表 3).

Table 2 Composition rules of product model

表 2 商品型号命名组成方式

命名方式	例子
英文字符串	Lumia, Galaxy, Ascend
数字字符串	3080, 118
英文与数字的混合字符串	700D, D5100, M35h
带有“-”符号的混合字符串	HC-V10GK, SB-910

Table 3 Naming rules of product model

表 3 商品名规则

品牌	商品名规则
佳能	(IXUS)(\s)*[\d]+(\s)*(HS|IS)*
尼康	(D|J)[\d]+(E|X|s)*[\s]*([\d]+-[\d]+[mm]*|[\d]+[mm]*)*
三星	(GX-|NX|WB|GC|MV|DV|EX|ES|ST|GN)[\d]+(S|F|M)*
HTC	(G|HD|Z|M|X|T)[\d]+[t|d]*

型号命名组成方式用于确定某字段是否可能是商品型号的组成部分,商品名规则用于确定某字段或字段

组合是否是型号.

3.4.1 关键字抽取

在抽取阶段,我们用以下两种方式从一个商品条目的商品描述抽取商品的品牌和型号,抽取的结果是关键字段的有序组合:

WE(word extraction):只抽取 Product List 中有的字段,并保持字段在商品描述中的前后顺序,这种方式认为商品描述只包含 Product List 中已有的商品.

WMAE(word match and extraction):根据商品命名组成方式匹配,将所有符合型号命名组成方式的字段提取出来,提取出的字段称为候选型号,同时提取出存在于 Brand List 中的品牌,将提取出的这些品牌称为候选品牌.保持候选型号和候选品牌在商品描述中的前后顺序.这种方式认为商品描述中可能包含除了 Product List 中已有商品外其他商品的信息.

经过抽取阶段后,一条商品条目的文字描述部分有两种字段组合.例如,有描述“佳能 A3300 IS/A3400 IS/A3500 IS 数码相机”,若 ProductList 中只能找到“佳能 A3300”,WE 的结果是“佳能”,“A3300”.根据 WMEA 抽取方法可得到 1 个候选品牌和 3 个候选型号的有序组合:“佳能”、“A3300”、“A3400”、“A3500”.若用两种抽取方式得到的字段组合相同,则将该抽取结果作为最终结果;若两种关键词有序组合不同,则需要找出较优的抽取结果.

3.4.2 找出最优抽取结果

给定多个品牌和型号的有序组合,我们认为字段之间关联性较大的组合为较优的结果.为此提出用隐马尔可夫模型 HMM 来解决此问题.如前文所述,构建一个完整的 HMM 首先需要确定模型的状态集以及观测集,然后根据状态集和观测集计算参数$\lambda=\{A,B,\pi\}$或者根据已有模型参数λ和观测序列 O 学习能使产生这一序列的概率最大的参数.本文提出的模型使用第 1 种方法构建模型.

候选型号是否为商品型号将直接影响 HMM 的构建.因此需要确定候选型号是否的确是商品型号,由于商品型号可能包括多个字段,故提出如下的识别方法:

WMEA 的结果是关键字段的有序组合,设条目的商品描述 D 的 WMEA 结果中共有 n 个候选型号 $w_1,w_2,\ldots,w_n$.可得到 n 个候选型号的从 1 元到 n 元的全排列组合 $P=\{P_{NA}\}$,$1\leqslant N\leqslant n$,$1\leqslant A\leqslant A_{NN}$,$N$ 表示 N 元,A 表示 N 元全排列组合的第 A 种情况.对于每个 P_{NA},若 P_{NA} 在 Product List 中已有,则忽略.若没有,则首先将 P_{NA} 用所有候选品牌的商品名规则去匹配,若 P_{NA} 不符合候选品牌的商品名规则,则到其余品牌的商品名规则中匹配,在以上匹配过程中,若 P_{NA} 符合某品牌 b 的商品命规则,则认为品牌 b 有商品型号 P_{NA},将品牌 b、型号 P_{NA} 以“B+M”形式添加进 Product List,并从 B2C 网站获取其价格,添加进 Product Price List,结束此 P_{NA} 的匹配工作.若所有被识别为商品型号的 P_{NA} 都不包含字段 w_k,$1\leqslant k\leqslant n$,则 w_k 不是任何商品的型号组成部分,将 w_k 添加进记录一般字段的文件 Not Model.例如,对于商品描述“2012 新款相机佳能 650d 18-55 套机 18-135 上海实体店”,WMAE 的结果是{“2012”,“佳能”,“650d”,“18-55”,“18-135”},在 Product List 中已经有“佳能 650d 18-55”,而“650d 18-135”符合“佳能”的商品名规则,所以“佳能 650d 18-135”添加进 Product List,而无论是“650d 18-55”还是“650d 18-135”都不包含字段“2012”,所以“2012”作为一般字段,添加进 Not Model.

将 Product List 中的所有字段作为状态,设状态集为 S,将 Product List 和 Not Model 中所有字段作为观测值,设观测值集为 V.HMM 过程如图 2 所示,其中,$q_i=s_n$,$1\leqslant i\leqslant t$,$1\leqslant n\leqslant|S|$,$s_n\in S$;$O_j=v_m$,$1\leqslant j\leqslant t$,$1\leqslant m\leqslant|V|$,$v_m\in V$.

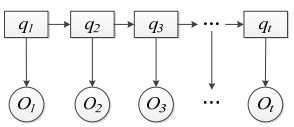


Fig.2 Illustration of HMM

图 2 隐马尔可夫模型过程图

HMM 的参数可由以下方法初始化:状态初始概率从 Product List 中计算得到.Product List 的记录是“*B*+*M*”形式,品牌字段和型号字段以及型号字段之间的前后关系代表了状态之间的转移关系,因此,状态之间的转移概率也根据 Product List 计算得到.考虑到商品之间往往有非常相似的名称,如尼康的相机名“D5100”、“D3100”,在状态“D5100”时,最有可能观测到的值是“D5100”,也有较大的概率观测到“D3100”,因此用状态与观测值之间的文本相似性 Levenshtein 距离(又称编辑距离,指的是两个字符串之间,由一个转换成另一个所需的最少编辑操作次数)[7]来度量在状态 X 时能够观测到观测值 Y 的概率[8].$HMM(A,B,\pi)$的参数设置如下:

$$A(x,y)=P(y\mid x)=\frac{count(x,y)}{\sum_z count(x,z)},\ x,y,z\in S \tag{1}$$

$$B(x,y)=P(y\mid x)=a\times Levenshtein(x,y),\ x\in S,\ y\in V \tag{2}$$

$$\pi(x)=P(x)=\frac{count(x)}{\sum_y count(y)},\ x,y\in S \tag{3}$$

其中,z 是 x 所有可转移到的状态,$count(x,y)$表示从状态 x 转移到状态 y 的情况出现的次数,$count(x)$表示状态 x 出现的次数,这里 a 用来正规化 B 的分布.

我们将 WE 和 WMEA 的结果中每个字段作为观测值,因此 WE 和 WMEA 的结果是观测值序列,用 HMM 的评估方法计算各种观测值序列出现的概率,概率大的选为最优的抽取结果.这里值得注意的是,根据评估方法的 Forward-backward 算法,观测值序列包含的观测值越多,计算出的概率肯定越小,因此,若 WE 和 WMEA 结果的字段数不同,设字段数少的抽取结果有 n 个字段,则对于字段数多的抽取结果,从中取出所有的 n 元词有序组合,将每种 n 元词有序组合看作不同的观测值序列,计算概率,用其中最大的概率作为字段数较多的抽取结果的概率.

3.5 误导性商品描述的识别

通过上述 HMM 过程,每个商品条目的商品描述可确定为一种关键字段的有序组合.用第 3.4 节描述的商品识别过程可得出该描述所包含的所有商品,但描述所指的商品应该是唯一的.从 Product Price List 中查询所有包含商品的价格,选择价格与条目价格最相近的商品为该条目的商品描述所指的商品.

对所有条目重复上述过程.可得到条目到商品的映射,将有相同映射的条目归到一组,称为一个商品组,在这样的组内,如果有误导性条目,则一定存在该商品的底价,而组内的误导性条目的价格低于该商品的底价.

我们利用统计上的四分位间距(inter-quartile range,简称 IQR)[9]识别异常值的思想来找出误导性条目.在统计学中,如果有一个完整的数据分布,设下四分位价格为 $Q1$,上四分位价格为 $Q3$,$IQR=Q3-Q1$,数据大于 $Q3+1.5\times IQR$ 或者小于 $Q1-1.5IQR$,则是异常值.

显然,相对于高价商品条目而言,低价商品条目更能吸引用户到其购买页面.因此,在商品组中,误导性条目都是价格较低的.因此我们只考虑低价的商品条目,商品的底价就是误导性条目价格的最高值.设误导性条目的价格的最高值为 max,设 $max=Q1-B\times IQR$,若商品组中条目的价格小于等于 max,则认为该条目为误导性条目.假设有误导性条目的商品组存在这样一个 B,通过训练集得到该值,用于预测新的商品组是否有误导性条目.

4 实 验

4.1 数据介绍

我们选用新蛋网(http://www.newegg.cn)和淘宝网(http://www.taobao.com)作为数据源.新蛋网是中国国内主营数码产品的 B2C 网站,数码产品在新蛋网上的价格可看做其在当前的标准价格.从新蛋网上爬取了相机、镜头、手机、电池等共 1 072 种电子产品的信息,包括每种产品的品牌、型号以及价格.用爬取的商品信息建立 Product List,同时建立与 Product List 对应的商品标准价格列表 Product Price List.本文只考虑有品牌有型号的正规产品,没有品牌或型号的产品此处不予考虑.

由于商品可用品牌字段和型号字段唯一确定,因此将 Product List 中的记录作为搜索词用淘宝网的搜索引

擎获取搜索结果.搜索结果的商品描述中都包含了搜索词.在实际数据收集过程中,从新蛋网获取商品信息与从淘宝网获取搜索结果是并行进行的,所以可以保证新蛋网上商品的价格与搜索结果中商品的价格是同一时间阶段的价格.在某搜索词的搜索结果中,价格过低的商品条目通常提供的是该商品的相关零配件,如手机壳、相机包等,因此我们将条目价格小于搜索词标准价格 50%的商品条目过滤掉.

在找出最优抽取结果过程中,新商品的价格同样从新蛋网获取.算法 1 的前两步一共识别出 1 153 种新商品,所以一共有 1072+1153=2225 种商品,对应了 2 225 个商品组,但由于新商品的组内条目数量都非常少,远远不能代表某种商品条目的完整分布,因此不作为有效的商品组.

在 1 072 种商品中,有 107 种商品属于相机,186 种商品属于手机,其余的是电子产品的附件产品如电池、镜头、存储卡等.手机类和相机类的每个商品组中都有误导性条目,而其余的配件类几乎没有误导性条目,因此用以下的实验找出手机和相机的误导性条目.实验数据集中的误导性条目由人工根据条目信息与商品购买页具体信息判别并标注.

4.2 实验结果

我们用五折交叉验证(5-fold cross validation),对相机类 107 个组,手机类共 186 个组分别求其通用 B 值.相机商品组和手机商品组都分为 5 份,用于 5 次验证实验,每次实验商品组数量信息见表 4.每次实验的训练数据中,每个商品组的 B 值都不相同,但分布情况比较相似,大部分在 0.5~1 之间.具体的分布情况见表 5.每次实验的 B 值、准确率和召回率(保留 3 位小数)见表 6.

Table 4 Number of instances in test and training group

表 4 每次实验中测试集与训练集的商品组数量

	相机类		手机类	
	测试数据	训练数据	测试数据	训练数据
1st fold	21	86	37	149
2nd fold	21	86	37	149
3rd fold	21	86	37	149
4th fold	22	85	37	149
5th fold	22	85	38	148

Table 5 Distributions of B values of training data

表 5 训练数据的 B 值在不同范围的分布情况

B 值的范围	1st fold (%)		2nd fold (%)		3rd fold (%)		4th fold (%)		5th fold (%)	
	相机类	手机类	相机类	手机类	相机类	手机类	相机类	手机类	相机类	手机类
0~0.5	3.48	8.72	4.65	10.07	3.48	8.72	2.35	9.40	4.70	8.78
0.5~1	56.98	69.80	58.14	67.11	63.95	67.79	62.39	68.46	67.06	65.54
1~1.5	36.05	18.12	34.88	18.12	29.07	18.79	32.94	17.45	25.58	21.62
1.5~2	3.48	2.01	2.33	3.36	3.48	3.36	2.35	2.68	2.35	2.02
2~3	0	1.34	0	1.34	0	1.34	0	2.01	0	2.02

Table 6 Experimental result

表 6 实验结果

	B		准确率(Precision)		召回率(Recall)	
	相机类	手机类	相机类	手机类	相机类	手机类
1st fold	0.780	0.750	0.866	0.847	0.868	0.898
2nd fold	0.780	0.780	0.934	0.905	0.923	0.875
3rd fold	0.750	0.780	0.878	0.879	0.977	0.861
4th fold	0.800	0.780	0.906	0.895	0.757	0.892
5th fold	0.760	0.790	0.871	0.910	0.921	0.864
平均	0.774	0.776	0.891	0.887	0.889	0.878

对于相机,B 取 5 次实验的平均值 0.774,用全部数据即 107 个组验证,得到 $Precision$=0.895,$Recall$=0.892.同样地,对于手机,B 取 5 次的平均值 0.776,用全部数据即 186 个组验证,得到 $Precision$=0.887,$Recall$=0.878.

综上,本文所述“商品描述分解与识别”过程的作用是将从淘宝网返回的搜索结果“分组”.“分组”过程可保

证每个商品组内的商品描述所指的商品一致,若不“分组”,则将某搜索词的搜索结果中的所有商品条目作为一个商品组,称为“搜索结果商品组”,这样的组内有一部分商品条目与搜索词所指商品不一致,会影响误导性商品描述的识别结果.

运用四分位间距识别异常值的方法,是因为各商品组内的误导性商品条目价格的最大值 max 与各组内价格分布的关系比 max 值与商品价格的关系更加密切.

我们进行了以下对比实验:一方面,用四分位间距识别异常值的方法找出相机类和手机类的“搜索结果商品组”组内的误导性商品描述,得到准确率和召回率,证明本文所提出的“商品描述分解与识别”过程的必要性;另一方面,我们仍然使用“分组”后的商品组,但 max 的值不通过 max=$Q1-B\times IQR$ 获取.设某商品组所属商品的标准价格为 SP(standard price),max=$SP\times P$,得到此实验结果的准确率与召回率,证明四分位间距识别异常值的方法更加有效.实验结果见表 7.我们将“分组”得到的商品组表示为 PG(product group),将未经“商品描述分解与识别”过程,直接得到的“搜索结果商品组”表示为 PGSR(product group from search result).

从表 7 的数据可知,同样在用 max=$Q1-B\times IQR$ 的方式获取 max 值的情况下,无论是相机类还是手机类,“分组”过程都在一定程度上提升了实验结果的准确率和召回率.而在商品组同样用 PG 的情况下,对两类商品,用 max=$Q1-B\times IQR$ 的方式获取 max 值比用 max=$SP\times P$ 的方式获取 max 值对实验结果的准确率和召回率均有不小的提升.因此,本文提出的“分组”过程是有必要的,并且用四分位间距识别异常值的方法较为有效.

Table 7 Comparison result

表 7 对比结果

	商品组	PG	PGSR	PG
	max	$Q1-B\times IQR$	$Q1-B\times IQR$	$SP\times P$
相机类	*B*/*P*	*B*=0.774	*B*=0.925	*P*=0.674
	准确率	0.895	0.732	0.652
	召回率	0.892	0.747	0.665
手机类	*B*/*P*	*B*=0.776	*B*=0.892	*P*=0.596
	准确率	0.887	0.691	0.680
	召回率	0.878	0.673	0.662

5 总 结

本文主要提出了一种能够较为有效地找出 C2C 网站中误导性条目的方法,其中分解识别商品条目的描述、确定条目描述指的商品以及从大量商品条目中找出误导性条目有较大的挑战.考虑到 C2C 网站上商品条目的具体特性,使用了基于统计的异常值识别方法.该方法具有一定的针对性且效果较好.本文提出的方法主要使用了概率模型 HMM 等统计方法,实验结果表明,本文提出的方法能够较好地解决上述问题.

References:

[1] Wu MM. The combination of B2C and C2C. Journal of Shaoguan University, 2008,10:95−99 (in Chinese with English abstract).

[2] Zheng H, Wu KW, Zhu QH. Detection of C2C reputation fraud activities based on neural network and SNA. Application Research of Computer, 2011,5:1882−1885 (in Chinese with English abstract).

[3] Zhu YC, Zhang W, Yu CH. Detection of feedback reputation fraud in Taobao using social network theory. In: Proc. of the 2011 Int'l Joint Conf. on Service Sciences. IEEE Press, 2011. 188−192.

[4] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: Identifying density-based local outliers. ACM Sigmod Record, 2000,29(2): 93−104.

[5] Ester M, Kriegel HP, Sander J, Xu XW. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of the KDD. 1996. 226−231.

[6] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. In: Proc. of the IEEE Ultrasonics Symp. IEEE Press, 1989. 257−296.

[7] Heeringa WJ. Measuring dialect pronunciation differences using Levenshtein distance [Ph.D. Thesis]. Diss. University Library Groningen, 2004.

[8] Pu KQ. Keyword query cleaning using hidden Markov models. In: Proc. of the KEYS 2009. ACM, 2009.

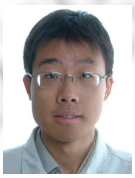
[9] Fisher RA. Statistical Methods for Research Workers. 5th ed., Edinburgh: Oliver and Boyd, 1934. 43−60.

附中文参考文献:

[1] 吴勉勉.论 B2C 与 C2C 两种模式的融合和发展趋势.韶关学院学报,2008,10:95−99.

[2] 郑华,吴克文,朱庆华.基于神经网络和 SNA 的 C2C 电子商务信誉欺诈识别研究.计算机应用研究,2011,5:1882−1885.

龙吟(1992－),男,重庆人,硕士生,主要研究领域为数据挖掘.
E-mail: rclongyin@126.com



胡鹤(1976－),男,博士,副教授,主要研究领域为人工智能,语义 Web.
E-mail: hehu@ruc.edu.cn

刘红岩(1968－),女,博士,教授,博士生导师,主要研究领域为数据挖掘,商务智能,社会计算.
E-mail: hyliu@tsinghua.edu.cn

杜小勇(1963－),男,博士,教授,博士生导师,主要研究领域为智能信息检索,高性能数据库实现,语义网技术.
E-mail: duyong@ruc.edu.cn

何军(1962－),男,博士,教授,博士生导师,主要研究领域为数据库,信息检索,数据挖掘.
E-mail: hejun@ruc.edu.cn