基于朴素贝叶斯学习的电子商务网站客户兴趣分类的应用研究*)

潘志方
(温州医学院计算机系　温州 325035)

摘　要　随着电子商务的不断发展,用户的分析和分类对电子商务网站来说越来越重要。因此需要一个行之有效的方法来进行用户分类并对其进行个性化服务。在本文中,我们提出了一种可以根据用户的网页访问记录和网上交易记录来动态地对顾客进行分类的方法,主要是利用了改进型的朴素贝叶斯分类器,对用户在网站上的行为进行分类,从而得到用户的分类信息,其结果可以作为提供个性化服务的依据。文章通过实验证明了上述方法的有效性和正确性。

关键词　朴素贝叶斯分类器,电子商务,用户分类

A Customer Classification Based on Naïve Bayes Classifier

PAN Zhi-Fang
(Department of Computer, Wenzhou Medical College, Wenzhou 325035)

Abstract　With the increasing interest and emphasis on customer demands in e-commerce, it is highly desired to extract customer features effectively and analyze customer orientations. This paper presents a new approach that employs a modified naive bayes classifier to group users dynamically based on their web access. Such a customer clustering method should be performed prior to internet stores as the basis to provide individual service. The experimental results of this clustering technique show the efficiency of the algorithm.

Keywords　E-commerce, Customer classification, Naïve Bayes

电子商务市场已成为发展最快的市场之一,从事电子商务营销的企业或网站已在世界经济领域崭露头角。社会上越来越多的企业商家,都认识到开展电子商务是在未来激烈的竞争中立于不败之地的重要保证,也是现时增加营业额的有效途径。网络营销已成为企业界的共识。大企业凭借着自身强大的实力建立和运作自己的网络。然而,在全面实施电子商务的过程中,面临着诸多问题。

1　客户关系管理

营销学术界以及企业界都认识到"保留客户,而不仅仅是获取客户"对于企业市场份额以及长期财务利益的战略价值。客户是企业生存和发展的基石,企业得到客户的信任,就留住了客户,赢得了市场。有效地进行客户分类,把握住不同客户的不同需求,才能更好地为客户服务,企业才能不被客户抛弃,顺利地发展壮大。

随着现代生产管理制度的建立和现代生产技术的快速发展,以产品为中心的模式正在向以客户为中心的商业模式改变,因此企业与客户之间必须建立一种适应新模式下的新型关系,在以客户为中心的商业模式中,如何以较低的成本最大限度地满足客户的个性化需求,已成为企业在市场竞争中处于优势的根本途径。要想快速地满足客户的个性化需求,企业必须具备快速准备地获取客户个性化需求的能力,并对客户需求进行分析、归类,进而把客户需求转换成企业的数据。

据欧美学者的研究表明,开发一个新客户是留住一个老客户成本的五、六倍,而20%的客户会为企业带来80%的收益。从这两方面看来,能够给企业带来持续收益的老客户是企业最宝贵的财富。因此,企业应对现有客户加以区分,发掘对企业极具价值的客户,对他们实施包括提高满意度等一系列关系营销方法,注重与这些客户建立长期良好的关系,留住他们以使他们成为企业忠诚的客户,从而使企业获得持续的高额收入并赢得竞争优势。可以通过对特定客户背景信息的分析,预测该客户所属的客户类别,从而采取相应的经营策略,这样可以有效地利用企业有限的资源,既能够提高企业的服务水平,开发客户资源,避免客户的流失;又能够节约企业资源,利用最小的投入,获得到较大的收益经营策略。

客户关系管理(Customer Relationship Management, CRM)是近几年来管理学界的热门话题。客户关系管理主张发展企业与客户的伙伴关系,并从中赢得利润;主张以客户、潜在客户的利益和需求,重组企业组织和工作流程;主张区别对待客户,针对客户实施目标营销,特别是高值客户的一对一营销,并从目标营销(尤其是一对一营销)中,建立区别于竞争对手的产品和服务,建立企业与客户之间的诚信。客户关系管理和相关的客户关系营销是企业营销的重要理论基础,它们有助于从公众营销转向定向营销,从粗放型的企业管理转向细致型经营。

随着Internet技术的发展和电子商务开始普及并走入普通消费者的生活,这为企业提供了一种全新的交易平台,在电子商务的新型服务模式中,企业可以通过与消费者之间的交流信息和他们的购物记录来管理特定消费者的网络行为。这种对消费者的了解可以将顾客信息转换为高质量的服务,还

*)浙江省高校青年教师资助计划研究课题。潘志方　讲师,硕士,主要研究方向:信息管理和系统,医学软件。

可以为产品的完善和改进提供具体依据。但是，面对数量巨大的消费者，企业如何才能知道他们的兴趣和爱好呢？这个问题的答案是建立个性化的服务模型。个性化也就是针对不同的人群提供不同的服务，企业可以通过顾客的购物行为和习惯来进行有针对性的市场划分和产品开发，以吸引不同类型的消费者，并提供他需要的信息产品和服务。通过这种方法可以提高消费者的满意度和忠诚度，并提高顾客的网上浏览频率，从而提高网上商品的销售额，最终使得电子商务企业受益。我们的研究目的就是要使用朴素贝叶斯分类器，根据客户的兴趣和偏好来对他们进行分类。这是在网络上为消费者提供个性化服务的基础。

2 朴素贝叶斯学习方法

贝叶斯推理提供了推理的一种概率手段。它基于如下的假定，即待考查的量遵循某概率分布，且可根据这些概率及已观察到的数据进行推理，以做出最优的决策。朴素贝叶斯学习算法能够计算显式的假设概率，是解决相应学习问题的最实际的方法之一。Michie 等[1]详细研究并比较了朴素贝叶斯分类器和其他学习算法，包括决策树和神经网络，发现朴素贝叶斯分类器在多数情况下与其他学习算法性能相当，在某些情况下还优于其他算法。特别是在文本分类的学习任务，它是最有效的算法之一。

朴素贝叶斯分类器基于一个简单的假定：在给定目标值时属性值之间条件相互独立，即在给定实例的目标值情况下，观察到联合的 $a_1, a_2, \cdots, a_n$ 的概率等于每个单独属性的概率乘积

$$P(a_1, a_2, \cdots, a_n | v_j) = \prod_i P(a_i | v_j) \quad (1)$$

贝叶斯方法的新实例分类目标是在给定描述实例的属性值 $\langle a_1, a_2, \cdots, a_n \rangle$ 下，得到最可能的目标值 v_{MAP}

$$v_{MAP} = \arg\max_{v_j \in V} p(v_j | a_1, a_2, \cdots, a_n) \quad (2)$$

(1)式代入(2)式中，可得到朴素贝叶斯分类器所使用的方法：

$$v_{NB} = \arg\max P(v_j) \prod_{\substack{i \\ v_j \in V}} P(a_i | v_j) \quad (3)$$

对于每个客户浏览过的网页和购买过的商品建立兴趣配置文件，每次客户登录时向他介绍感兴趣的商品。由于兴趣配置文件以文本为形式，这样，朴素贝叶斯分类器就可以应用于兴趣配置文件。朴素分类器是目前所知文本文档分类算法中最有效的方法之一，可以得到很好的效果。

3 用于学习和分类文本的朴素贝叶斯算法

以下两个过程，其中 LEARN_NAIVE_BAYES_TEXT 用来分析所有训练文档，抽取所有出现的英文单词、中文字、记号，然后在不同目标类中计算其频率以获得必要的概率估计。当有了一个待分类的新实例，过程 CLASSIFY_NAIVE_BAYES_TEXT 使用此概率估计来计算 v_{NB}。

w_k 代表词典中的第 K 个字，n 为所有目标值为 v_j 的训练样例中单词位置的总数，n_k 是在 n 个单词位置中找到 w_k 的次数，而 $|Vocabulary|$ 为训练数据中的不同英文单词或中文字(以及记号)的总数。

LEARN_NAIVE_BAYES_TEXT(Examples, V)

Examples 为一组文本文档以及它们的目标值。V 为所有可能目标值的集合。此函数作用是学习概率项 $P(w_k | v_j)$，它描述了从类别 v_j 中的一个文档中随机抽取的一个词(中文或英文)为 w_k 的概率。该函数也是学习类别的先验概率 $P(v_j)$。

(1)收集 Examples 中所有的词、标点以及其他记号

• Vocabulary←在 Examples 中任意文本文档中出现的所有词及记号的集合

(2)计算所需要的概率项 $P(v_j)$ 和 $P(w_k | v_j)$

对 V 中每个目标值 v_j

• $docs_j$←Examples 中目标值为 v_j 的文档子集

• $P(v_j) \leftarrow \dfrac{|docs_j|}{|Examples|}$

• $Text_j$←将 $docs_j$ 中所有成员连接起来建立的单个文档

• n←在 $Text_j$ 中不同词位置的总数

• 对 Vocabulary 中每个词 w_k

n_k←词 w_k 出现在 $Text_j$ 中的次数

$$P(w_k | v_j) \leftarrow \frac{n_k + 1}{n + |Vocabulary|}$$

CLASSIFY_NAIVE_BAYES_TEXT(Doc)

对文档 Doc 返回其估计的目标值。A_i 代表在 Doc 中的第 I 个位置上出现的词。

• positions←在 Doc 中的所有词位置，它包含能在 Vocabulary 中找到的记号

• 返回 v_{NB}

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i \in position} P(a_i | v_j)$$

汉字由于其特殊性汉字分词可采用一些成熟的自动分词系统，或就一个字一个字地进行分割。购买的商品就以其货品名作为分词依据。属性分割时，兴趣配置文件中的货品名优先分词，客户浏览过的其他货品名、非货品名的汉字以一个字进行分词，英文按单词进行分词。

这样就可以开始应用朴素贝叶斯分类器了。解决两个重要的设计问题：一是如何估计朴素贝叶斯分类器所需的概率，二是怎样将任意文档表示为属性值的形式。前者可以按上面的算法进行解决。关于后者，可对每个此的位置定义一个属性，该属性的值为在此位置上的词。如电子商务网站中有一个文本：感应洁具一全自动感应洗手器，感应龙头。

这样上例中的文本被表示为 16 个属性，对应 16 个字的位置。第一个属性值为“感”，第二个为“应”，依次类推。很显然，较长文档的属性数目也是较多的。在应用时，先用训练文档进行训练，训练文档中分类为 *like* 的文档都来源于客户的兴趣配置文件，分类为 *dislike* 的文档可来源于网站中客户从不浏览的网页内容。这样，如果有了一个新文档要分类判断是否属于向客户推荐内容，就可以应用朴素分类器了。比如新文档就是上述例子的话，那么应用分类器如下：

$$\begin{aligned} v_{NB} &= \arg\max_{v_j \in \{like, dislike\}} P(v_j) \prod_{i=1}^{16} P(a_i | v_j) \\ &= \arg\max_{v_j \in \{like, dislike\}} P(v_j) P(a_1 = \text{“感”} | v_j) P(a_2 = \text{“应”} | v_j) \\ &\quad \cdots P(a_{16} = \text{“头”} | v_j) \end{aligned}$$

这里需要两个独立性假定：一是在一个位置上出现某英文单词或中文字的概率独立于另外一个位置的单词或字；二是假定可遇到一个特定单词 w_k 的概率独立于单词所在位置。

4 应用结果

这种改进的混合了英文单词和中文字的朴素贝叶斯算法

(下转第 222 页)

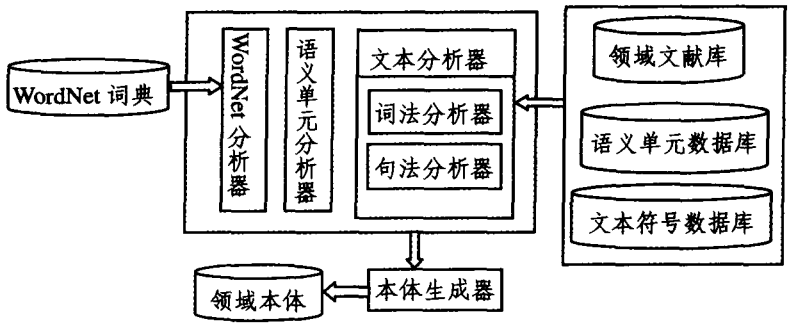

图 4　本体构建系统总体框架

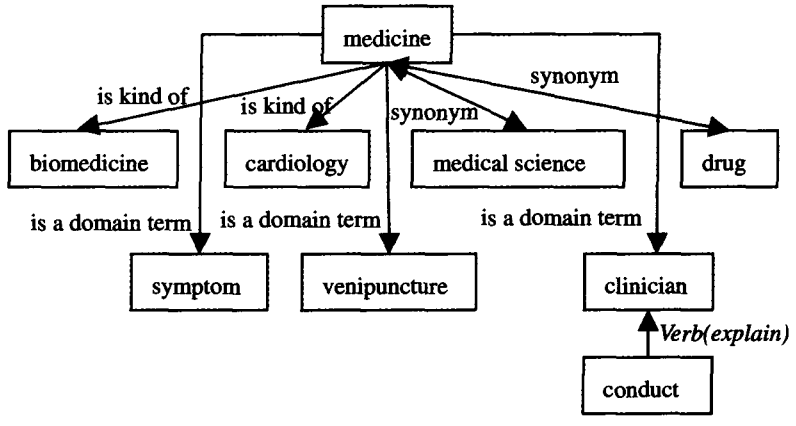

图 5　目标本体片断

表 1　目标领域本体概要

概念类型	数量	所涉及到的关系	关系数量
同义词概念	2	通过同义关系添加到本体中	2
上下位概念	117	通过概念之间的上下位关系添加到本体中	117
整体部分概念	47	通过政体部分关系添加到本体中	47
领域术语概念	373	通过隶属领域关系添加到本体中	373
文本挖掘概念	78	通过动词关系添加到本体中	457

结束语　使用通用本体 WordNet 与自然语言处理技术相结合的方法构建本体，能够在领域核心本体的原型基础上，进行本体的提炼，这是一种快速的本体构建方法。使用这种方法向领域本体中添加更多的概念是一个迭代的过程，能够根据本体构建者的要求，很好地进行过程控制。基于语义单元中动词的概念间关系的确定，对于领域本体中概念之间关系的扩展具有重要的意义。

由于语义单元反映的是一个简单句子片段的语义属性，在以后的研究中，如果能够通过统计方法，对语义单元之间的语义关联性进行挖掘，使更多的语义单元进行合理组合，形成更强的语义属性，将使本体的构建无论从质量上和效率上都会大大提高，这也是本文所描述的本体自动构建方法的重点改进方向。

参 考 文 献

1　邓志鸿，唐世渭，张铭，杨冬青，陈捷．ontology 研究总述．北京大学学报，2002，38(5)

2　Chandrasekaran B, Josephson J R, Benjamins V R. What Are Ontologies, and Why Do We Need Them? IEEE Intelligent Systems, 1999, 14(1): 20～26

3　Noy N F, McGuiness D L. Ontology Develo- pment 101: A Guide to Creating Your First Ontology. SMI technical report SMI-2001-0880 (2001), Stanford University

4　WordNet. a lexical database for the English langu- age. http://wordnet. princeton. edu/

5　GATE. A General Architecture for Text Engineer- ing. http://gate. ac. uk/

6　The Stanford Parser. http://nlp. stanford. edu/softwa re/lex-parser. shtml

(上接第 215 页)

效率如何呢？我们用 Java 构建了一个电子商务平台，在 20 个人中进行模拟应用，收集了他们的浏览喜好和购物记录形成了兴趣文件，并将数据进行分析，用上述方法进行了用户分类，当用户上网站时进行货品推荐，均达到了比较高的相关度和满意度。

结束语　正确识别客户是客户关系管理成功的基础，也是电子商务网站对客户进行有效管理的前提条件。应该对客户进行正确的分类，明确各类客户之间的关系以及各类客户管理工作之间的相互联系，以便更好地留住客户，产生更大的效益。

参 考 文 献

1　Michie D, Spiegelhalter D J, Taylor C C. Machine learning, neural and statistical classification[M]. New York: Ellis Horwood, 1994

2　曹渝昆，李云峰，汪成亮，等. 改进型模糊神经网络在顾客分类中的应用研究[M]. 计算机工程与应用，2006，19：218～221

3　Mitchell T M. 机器学习[M]. 曾华军，等译. 北京：机械工业出版社，2003