请注明引用地址:从用户评论中挖掘产品属性*——基于SOM的实现[J],现代图书情报技术,2009(5):61-66.

从用户评论中挖掘产品属性*
——基于SOM的实现

余传明
(上海理工大学管理学院 上海 200093)

【摘要】在分析现有产品属性识别方法不足的基础上,提出一种利用自组织映射(SOM)进行属性识别的新方法,定义一种新的名为"属性叠加矩阵"的SOM显示方式。为验证该方法的有效性,以餐馆评论为样本,从中抽取饮食行业的产品属性。实验证明提出的方法识别产品属性的效果较好。
【关键词】观点挖掘 属性识别 自组织映射 评论 属性叠加矩阵
【分类号】TP183 TP391

Mining Product Aspects from User Reviews
——An SOM - based Approach

Yu Chuanming
(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】This paper first analyzes the limitation of the existing methods of aspect identification. Then a novel method is presented which utilizes Self - organization map to identify the aspects from product reviews. A new SOM display named "Attribute Accumulative Matrix" is defined. In order to verify the validity of the method, we extract the product aspects from the restaurant reviews on a website. The experiment results show that this approach can effectively extract the product aspects.
【Keywords】Opinion mining Aspect identification Self - organization map Reviews Attribute accumulative matrix

1 前 言

随着Web 2.0的迅速发展,互联网逐渐成为人们表达观点、情感的重要渠道。互联网上的主观信息迅速增长,对这些观点和评论进行分析能够帮助企业改进产品、提高质量,并及时修复潜在可能恶化的客户关系,因而具有非常重要的意义。

在这种情况下,越来越多的公司把目光投向互联网上的产品评论,开始分析这些评论中所传递的重要信息。然而,由于产品评论数量巨大且呈现无结构化的特点,通过人工阅读的方式难以完成。如何解决海量的评论信息与有限的人工阅读能力之间的矛盾,成为摆在研究者面前的一个重要问题。

为了解决上述问题,研究者开始考虑使用自动化的方式对网上评论进行分析。在早期的研究中,学者倾向于判断整篇评论的极性,即研究评论者在整篇评论中所体现的对于某个产品的态度[1-3]。但是,对整篇评论给出极性判断对于企业了解用户反馈并不具备太多的实际意义。评论者可能对于产品的某些属性给予正面评价,而对

收稿日期:2009-02-04
收修改稿日期:2009-02-24
*本文系国家自然科学基金项目"基于随机服务理论的复杂网络和人类动力学演化模型"(项目编号:70871082)、上海市第三期重点学科建设项目"管理科学与工程"(项目编号:S30504)和上海市第三期本科教育高地建设项目(电子商务)的研究成果之一。

于另外一些属性却给予负面评价。如果不了解这些负面评价所针对的产品属性,则很难帮助企业提升产品竞争能力。

越来越多的研究开始建立在产品属性识别的基础上,即在大规模产品评论中,首先识别产品属性,再分析带有主观评价的句子中所传递的对于产品的某项属性的态度。

2 产品属性识别的相关方法

产品属性多表现为名词和名词短语,因此属性识别一般是对被评论的名词和名词短语进行识别。从实现方法上看,可以分为人工定义和自动识别两种思路。

(1)人工定义就是针对特定领域的产品建立该领域的属性词汇表。Zhang L[4] 等人利用人工定义针对电影的产品属性,将电影的产品属性分为电影元素(如 Screenplay, Vision Effect)以及与电影相关的人员(如 Director, Screenwriter, Actor)。如果采用人工定义产品属性的方法,那么每一个领域的产品都需要有该领域的专家参与才能定义该领域的产品属性。因此,这种方法可移植性较低,一旦产品的功能发生改变,则需要重新确定。

(2)自动识别产品属性则主要采用词性标注、句法分析等自然语言处理技术对产品评论中的语句进行分析,从中识别产品属性。在 Hu M 等人的研究[5]中,他们将词性标注后的句子中的名词和名词短语抽取出来,置于集合 D,基于 Apiori 算法得到候选频繁特征集合(Candidate Frequent Features),利用紧凑剪枝法(Compactness Pruning)去掉没有意义的多词短语,利用冗余剪枝法(Redundancy Pruning)去掉冗余的单个词构成的词汇。该方法的优点在于算法简单,易于统计,缺点在于可能产生大量的冗余。在 Kim S M 等人的研究[6]中,他们首先寻找句子中包含主观极性的词汇,然后定义一个大小固定的窗口,以主观性词汇为中心,将窗口中的名词或者名词短语作为属性。这种方法能够避免大量冗余名词的产生,但需要以极性词汇表作为基础,限制了该方法的使用。Popescu A M 等人[7]改进了 Hu M 的方法,他们通过计算名词或者名词短语与一定的区分符之间的共现程度来找出产品属性,共现的度量值利用点互信息 (Point - Wise Mutual Information)来衡量。该方法克服了产生大量冗余的问题,但是区分符的统计往往需要利用词典或者加入较多的人工因素,这给研究带来一定的难度。

使用机器学习的方法进行属性识别目前也是一个热点,包括概率潜在语义分析法(Probabilistic Latent Semantic Analysis)[8]、潜在狄利克雷分布法(Latent Dirichlet Allocation)[9]、相关主题模型法(Correlated Topic Model)[10]等。概率潜在语义分析法往往按照品牌而不是属性对产品进行分面,因此属性识别效果并不佳。此外,由于参数的个数随着语料库中文档数量的增加而线性增加,这可能会造成过度拟合,即所学习到的模型无法应用到新的文档。潜在狄利克雷分布法减少了参数的个数,也降低了人工参与的成分。但是,它同样是按照品牌而不是属性进行分面,因此属性识别效果也不好。相关主题模型法是对潜在狄利克雷分布法的扩展。在潜在狄利克雷分布法中,相关潜在主题的分布是基于狄利克雷分布进行计算。而在相关主题模型法中,狄利克雷分布(Dirichlet Distribution)被逻辑正态分布(Logistic Normal Distribution)所取代,逻辑正态分布的协方差矩阵用来表示潜在主题之间的关联。由于考虑相关主题的关联而引入的逻辑正态分布增加了模型的精确度。但是,由于它与多项分布具有非共轭性,这导致模型的使用具有较大的难度。

本文在分析现有的产品属性识别方法不足之处的基础上,提出一种利用自组织映射的无监督学习进行属性识别的新方法。为了验证该方法的有效性,笔者以餐馆评论为例,从中抽取饮食行业的产品属性。

3 基于自组织映射的产品属性识别

3.1 自组织映射

自组织映射(Self - Organization Map)是由芬兰赫尔辛基大学 Teuvo Kohonen 教授提出的一种两层神经网络[11]。它由输入层和竞争层组成,输入层是一维的神经元,竞争层为一维或二维的分类空间。在 SOM 模型中,每一个权向量 $W_i = (w_{1i}, w_{2i} \cdots, w_{mi})$ 都可以看作是神经网络的一种内部表示,它是与输入序列 $X = (x_1, x_2, \cdots x_n)$ 相对应的有序映象。SOM 训练过程由网络中权向量的自组织过程和最优匹配神经元的选择两部分组成,具有保留输入数据的拓扑结构的特点,能够将高维数据映射到低维空间中,属性相似的输入样本会映射到 SOM 空间的相近位置。

3.2　自定义的属性叠加矩阵及其原理

在 SOM 输出中,属性相似的输入样本会映射到 SOM 空间的相近位置,典型的输出形式为统一距离矩阵(即 U - matrix)[12],它通过计算相邻节点的权向量之间的欧几里德距离,显示各个样本的聚类及其边缘。但是用户无法从这些聚类中看出输入样本的属性叠加效应。因此,本文提出一种名为"属性叠加矩阵"的新概念,并应用于 SOM 输出的背景颜色,其定义如下。

设某 SOM 输出有 u 行 v 列,如等式(1)所示:

$$S=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1v} \\ s_{21} & s_{22} & \cdots & s_{2v} \\ \cdots & & & \\ s_{u1} & s_{u2} & \cdots & s_{uv} \end{pmatrix} \quad (1)$$

其中 s_{ij}($i=1,2,\cdots,u$, $j=1,2,\cdots,v$)表示与第 i 行、第 j 列的 SOM 结点相联系的权向量,包含 n 个分量,分别表示为($w_{ij1}, w_{ij2}, \cdots, w_{ijn}$),其中 n 为输入样本的属性个数,即维度。

属性叠加矩阵 C 也有 u 行 v 列,如等式(2)所示:

$$C=\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1v} \\ c_{21} & c_{22} & \cdots & c_{2v} \\ \cdots & & & \\ c_{u1} & c_{u2} & \cdots & c_{uv} \end{pmatrix} \quad (2)$$

其中 c_{ij} 为属性叠加矩阵中第 i 行第 j 列的元素($i=1,2,\cdots,u$, $j=1,2,\cdots,v$)。首先,计算 c'_{ij} 的值,其计算方法如等式(3)所示:

$$c'_{ij}=\sum_{k=1}^{n} w_{ijk} \quad (3)$$

然后计算属性叠加矩阵中的元素 c_{ij},如等式(4)所示:

$$c_{ij}=\sqrt{\frac{c'_{ij}}{\max(c'_{ij})}} \quad (4)$$

从(3)、(4)式中可以看出,属性叠加矩阵中的每个元素的值等于对应结点的权向量的所有分量之和除以其出现的最大值的商的平方根。将属性叠加矩阵中所有元素的值转换为不同的颜色,并应用于对应的 SOM 结点的背景颜色,研究者可以根据 SOM 结点的背景颜色判断属性叠加矩阵中元素值的大小。在(4)式中之所以要将 c'_{ij} 除以其最大值的商开平方是为了拉开属性叠加矩阵中各元素的值的差距,避免将属性叠加矩阵应用于 SOM 输出的背景时,大多数结点的背景颜色过于一致,使用户无法区别输入样本的属性叠加效应之间差别。

3.3　基于属性叠加矩阵的产品属性识别

在产品属性识别研究中,以名词(短语)为行,以评论为列,构造输入矩阵。输入矩阵中某个元素的值等于该元素所在的行代表的名词(短语)在对应的列代表的评价中出现的次数。利用 SOM 算法对该矩阵进行训练。由于 SOM 具有保留输入数据的拓扑结构的特点,出现在相似的评论集合中的名词(短语)将映射到相邻的 SOM 结点中,因此 SOM 输出将按照名词(短语)所出现的评论集合对名词(短语)进行聚类。根据属性叠加矩阵的定义,属性叠加矩阵中某个元素的值反映了映射到该结点的名词(短语)在所有评论中出现的次数之和,如果该元素的值较大,则表示映射到该结点的名词(短语)在评论中出现次数较多,那么它成为产品属性的可能性越大。

设想评论中可能存在某些名词(短语)并不是产品属性,可能是与客户自身相关的事物,例如"今天我去××餐馆过生日",其中名词"生日"显然不是产品属性。然而这种非产品属性的名词(短语)在评论中出现的情况较为分散,利用 SOM 可以将名词(短语)按照其所出现的评论集合进行聚类,同时属性叠加矩阵能够反映其在评论中出现的次数,即各个名词(短语)的热门程度,这样就可以筛选出真正的产品属性。

为了验证属性叠加矩阵对于属性识别的有效性,笔者以餐饮业为例,从网上收集了相关评论,在进行分词和词性标注后,利用 SOM 和自定义的属性叠加矩阵进行属性识别的实验。

4　实验过程及结果分析

4.1　网络数据收集

点评网(http://www.dianping.com)是国内著名的评论网站,该网站包括数十万家注册餐馆(其中上海地区 31 394 家),数百万个注册用户和数千万条评论数据。笔者首先从点评网上抽取了餐馆数据(包括餐馆名称、地址、别名等),然后对每一家餐馆,抽取相应的用户评论信息。这里选择了评论最多的上海地区的 300 家餐馆和相应的 22 157 条评论作为研究对象。

4.2　分词与词性标注

针对每一篇评论,笔者使用中国科学院计算技术研究所的 ICTCLAS 分词系统[13]进行分词和词性标注,每篇评论标注后的格式如下所示:

“因为/c 是/v 家庭/n 聚餐/vn,/wd 所以/c 没有/d 吃/v 鲍鱼/n 鱼翅/n 之类/rz 的/ude1 菜/n,/wd 选/v 了/ule 比较/d 简单/a 的/ude1 菜/n,/wd 但是/c 味道/n 很/d 好/a 。/wj 因为/c 是/vshi 按照/P 老式/b 里弄/n 房子/n 建造/v 的/ude1 新房/n,/wd 所以/c 环境/n 很/d 不错/a,/wd 也/d 很/d 安静/a,/wd 包房/n 很/d 舒服/a 。/wj 从/P 环境/n 、/wn 菜式/n 、/wn 口味/n 等/udeng 方面/n 都/d 很/d 精致/a 、/wn 雅致/a,/wd 也/d 很/d 适合/v 上海/ns 人/n 的/ude1 口味/n 。/wj 红烧肉/n 非常/d 好吃/a,/wd 肥瘦/n 适中/a,/wd 非常/d 入味/v;/wf 黑椒/ng 牛肉/n 粒/ng 非常/d 的/ude1 嫩/a,/wd 连/ulian 我/rr 90/m 岁/qt 的/ude1 外婆/n 都/d 很/d 喜欢/vi 。/wj ”。其中,“/”后的字母为对应词汇的词性标注。

4.3 SOM 输入矩阵的构造

产品属性大多表现为名词和名词短语,因此在词性标注之后,只需要统计名词和名词短语的词频。为了提高 SOM 训练的效率,通过设置频度阀值,过滤掉总频度低于 6 次以及所出现的评论数少于 10 篇的词汇,得到由 3 671 个词汇所构成的候选产品属性集合。通过统计 3 671 个名词或短语所构成的候选产品属性集合在 2 000 篇评论中的分布情况,构建一个 3 671 行与 2 000 列的矩阵 M,其中元素 m_{ij} 代表第 i 个候选产品属性在第 j 篇评论中出现的次数。

4.4 SOM 训练

在这里使用的是赫尔辛基大学信息与计算机科学实验室开发的免费软件 SOM Toolbox[14],软件的使用基于 Matlab 环境,训练的过程分为归一化数据、初始化权向量、选择学习算法、输出结果等几个步骤。

(1)归一化数据

由于不同评论中各候选产品属性的词频数相差很大,为了避免数值范围较大的属性可能会在 SOM 显示中占统治地位,对第 4.3 节得到的矩阵的每一行做归一化处理。在这里使用了方差规整方法(Variance Normalization)[15],即归一化过程是线性的,所有属性的方差被规范化为 1。

(2)初始化权向量

关于权向量的初始化,有两种方法[16]:随机初始化与线性初始化。研究表明,在多数情况下,线性初始化优于随机初始化[16]。这里使用了线性初始化的方法,首先计算每个候选产品属性的特征值和特征向量,然后沿着若干个最大的特征向量初始化权向量。

(3)选择学习算法

SOM 学习算法有两种[16]:序列学习(Sequential Learning)算法和批学习(Batch Learning)算法。两者之间的重要区别在于,序列学习算法是每处理完一个输入数据就更新与该输入数据对应的获胜结点及相邻结点相联系的权向量,而对于批学习算法,更新与获胜结点及相邻结点相联系的权向量是等到所有的输入数据都处理完之后才进行,它将匹配到该结点及其相邻结点的所有数据向量的平均值设置为结点新的权向量。由于批学习算法具有速度快、无需调整适应参数以及结果可再现等优点[17],在这里选择了批学习算法进行训练,过程如下[18]:

①提取所有的输入样本,计算每个输入样本 s_i 与每个输出结点相联系的权向量(w_j)之间的欧几里德距离,选择最佳匹配节点 j^*,使输入样本与该结点相联系的权向量之间的欧几里德距离最小,即将所有输入样本映射到对应的输出结点中;

②对于每个输出结点 i,收集所有映射到结点 i 附近属于拓扑结构邻居集 N_i 中的输入样本;

③更新所有的输出结点 i 的权向量,新的权向量等于该结点及其相邻结点的权向量的平均值;

④重复第② - ③步直至收敛。

其中,邻居集 N_i 大小的定义与序列学习算法中邻居距离的定义是相似的。N_i 的逐渐减少意味着当重复第② - ③步时邻居的范围在缩小。在最后一次迭代时,N_i 可能只包括输出结点 i 本身。

(4)输出结果

为了避免平面输出产生的边缘效应[16],采用超环面[12]的 SOM 输出形状,应用第 3.2 节提出的属性叠加矩阵作为 SOM 输出的背景颜色,SOM 输出的结果如图 1 所示。它共有 312 个结点,每个结点中的数字代表映射到该结点的名词(短语)数量。右方的颜色条指示出 SOM 输出的背景颜色代表的属性叠加矩阵值的大小,例如红色代表属性叠加矩阵值较大,而蓝色代表属性叠加矩阵值较小。

4.5 SOM 的输出分析

图 1 显示,位于 SOM 输出的 4 个“角”(实际上在超环面空间中,这些角落与边缘是连在一起的)的少数几个结点的属性叠加矩阵的值较大,呈现红、橙、黄色,

表示映射到这些区域的名词(短语)在评论中出现的最为广泛;位于 SOM 输出的“中间偏下方”区域的属性叠加矩阵的值居中,呈现绿、青、淡蓝色,表示映射到这些区域的名词(短语)在评论中出现的较为广泛;而位于 SOM 输出的“上方中间”的大片区域的属性叠加矩阵的值较小,呈现深蓝色,表示映射到这些区域的名词(短语)在评论中出现的较少。

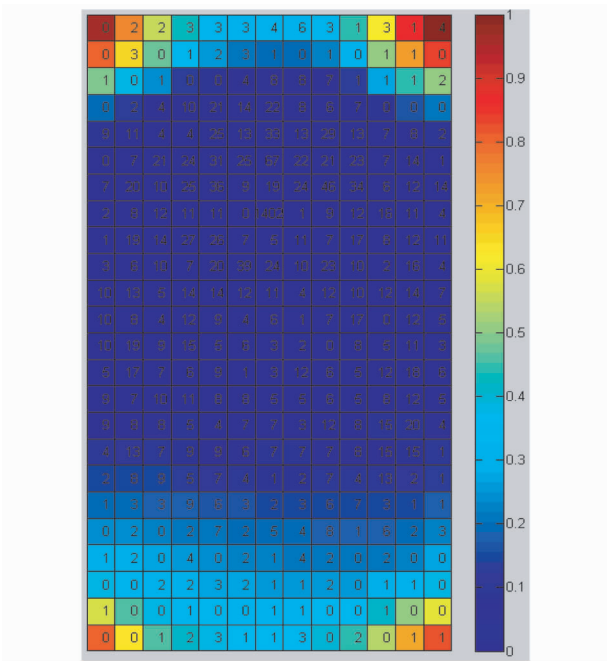

图 1　SOM 输出(以属性叠加矩阵为背景颜色)

根据图 1 中各结点的背景颜色,参照图右方的颜色条,将候选产品属性的热门程度分为 5 个等级,用数字 1 -5 来表示候选产品属性的热门程度,其中 1 表示最热门的候选产品属性,5 表示最冷门的候选产品属性,如表 1 所示(仅考虑非空结点)。其中结点 s(i,j)表示第 i 行、第 j 列的 SOM 结点。

根据表 1,候选产品属性可以按照其热门程度分为三类:

(1)热门程度为 1(即颜色为红色)的候选产品属性,为热点候选产品属性;

(2)热门程度为 2 -4 (即颜色为橙色、黄色、绿色、青色、淡蓝色)的主题,为次热点候选产品属性;

(3)热门程度为 5(即颜色为深蓝色)的主题,为一般性候选产品属性。

为了揭示所有候选产品属性在各个热门程度等级中的分布,将各个热门程度等级包含的候选产品属性数量以及热点、次热点和一般性候选产品属性的数量占全部主题数量的比例归纳在表 1 的右侧。计算结果显示,热门程度为 1 的候选产品属性数量最少,为 6 个,仅占全部候选产品属性的 0.16%;热门程度为 2—4 的候选产品属性数量也较少,为 390 个,占全部候选产品属性数量的 10.62%;而热门程度为 5 的候选产品属性数量最多,为 3 275 个,占全部主题数量的 89.21%。由此可见,随着候选产品属性热门程度的降低,对应的数量会增多。这说明,用户的评论热点主要集中在少数几个候选产品属性上,而大多数候选产品属性在用户评论中出现的次数较少,符合常见的“二八原则”。

表 1 候选产品属性及其热门程度

背景颜色	频繁等级	非空 SOM 结点	候选产品属性	数量	比例
红色	1	s(1,13),s(1,12),s(24,12)	服务 菜 环境 味道 口味 价格	6	0.16%
橙色	2	s(1,2),s(2,12),s(24,12)	感觉 地方 鱼翅	3	0.08%
黄色	3	s(2,2),s(1,11)	东西 套餐 时候 特色 量 餐厅	3	0.08%
绿色、青色、淡蓝色	4	s(1,4),s(1,10),s(2,11),s(3,1),s(3,10)等	店 老板 朋友 酒 商务 鹅肝 客户 鲍鱼 厨师 等	384	10.46%
深蓝色	5	s(4,4),s(4,5),s(4,6),s(5,1),s(5,2)等	北极 豆沙 和平 火候 土豆 觅女孩子 清汤 酒精 红茶 贝 五花肉 金属 捞饭 玻璃杯 鲨鱼 电影 工作餐 孩子等	3275	89.21%

从热点候选产品属性的内容来看,热门程度为 1 的候选产品属性为‘服务’、‘菜’、‘环境’、‘味道’、‘口味’、‘价格’。这些候选产品属性较好地反映了现实中用户最关心的产品属性,与 Google 公司的 Goldensohn 等人使用人工定义方式在静态属性抽取[19]中所确定的 4 个属性(即 Service,Food, Decor, Value)相一致,这说明本文所采取的方法能够较好地识别与被评论对象相关的产品属性。因此,可以选择热门程度为 1 的候选产品属性作为整个饮食行业的产品属性。热门程度为 2 的候选产品属性集中在‘感觉’、‘地方’、‘鱼翅’。‘感觉’和‘地方’两个词的出现说明用户评价对产品的属性定位具有一定的模糊性。在评论中会出现类似于“总的感觉还不错”、“这个地方比较不错”的语句,从这些语句中得出其所评论的产品属性具有一定难度。热门程度为 3、4、5 的候选产品属性呈

逐渐发散趋势,出现了‘鲍鱼’、‘红茶’、‘五花肉’等食品名称。

由上可知,产品属性被较好地集中在红色区域,这说明本文所提出的方法具有较好的产品属性识别效果。

5 结 语

本文提出了一种无监督学习的机器学习方法,即自组织映射人工神经网络方法,并定义了一种新的名为“属性叠加矩阵”的自组织映射显示方式。使用该方法能够在不需要人工干预的条件下从客户评论中识别产品属性,避免了领域专家参与属性定义所带来的可移植性问题。与其他的机器学习方法相比,它具有较好的去除冗余的效果,同时又不需要以极性词汇表或区分符作为前提,因而在实践中具有较强的实用性。

值得说明的是,对产品评论进行观点挖掘目前是一个刚刚出现的领域,相关研究在国外才刚刚起步,在国内则更少,有许多新的理论、方法和应用还需要探索和发现。如何在属性识别的基础上,对用户的态度进行自动分析,并给出自动化的归纳,是下一步研究的重点。

参考文献:

[1] Somasundaran S, Ruppenhofer J, Wiebe J. Detecting Arguing and Sentiment in Meetings[C]. *In: Proceedings of Workshop on Discourse and Dialogue*(*SIGdial*'2007), Antwerp, Belgium, September 2007:311 – 319.

[2] Yang C, Lin K, and Chen H H. Emotion Classification Using Web Blog Corpora[C]. In: *Proceedings of the* 2007 *IEEE/WIC/ACM International Conference on Web Intelligence*(*WI – IAT*'2007), Silicon Valley, U. S. A. 2007:275 – 279.

[3] Fung G P, Yu J X, Lu H. The Predicting Power of Textual Information on Financial Markets[J]. *IEEE Intelligent Informatics Bulletin*, 2005,5(1):1 – 10.

[4] Zhuang L, et al. Movie Review Mining and Summarization[C]. In: *Proceedings of ACM International Conference on Information and Knowledge Management*(*CIKM*'2006), Arlington, Virginia, U. S. A. 2006:1 – 7.

[5] Hu M, Liu B. Mining and Summarizing Customer Reviews[C]. In: *Proceeding of the* 10th *Knowledge Discovery and Data Mining Conference*(*KDD*'2004), Seattle, WA, U. S. A. 2004:168 – 177.

[6] Kim S. M, et al. Determining the Sentiment of Opinions[C]. In: *Proceedings of the* 20th *International Conference on Computational Linguistics*, Geneva, Switzerland. 2004:1 – 8.

[7] Popescu A M, Etzioni O. Extracting Product Features and Opinions from Reviews[C]. In: *Proceedings of Empirical Methods in Natural Language Processing*(*EMNLP*'2005), Vancouver, B. C., Canada. 2005:1 – 8.

[8] Hofmann T. Probabilistic Latent Semantic Indexing[C]. In: *Proceedings of the* 22nd *Annual International SIGIR Conference on Research and Development in Information Retrieval*, California, U. S. A. 1999: 1 – 8.

[9] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. *Journal of Machine Learning Research*, 2003(3):993 – 1022.

[10] Blei D. and Lafferty J. Correlated Topic Models[C]. In: *Proceeding of the* 20th *Annual Conference on Neural Information Processing Systems*, Vancouver, B. C., Canada. 2006:1 – 8.

[11] Kohonen T. Self – Organized Formation of Topologically Correct Teaturc Maps[J]. *Biological Cybernetics*, 1982,43(1):59 – 69.

[12] Ultsch A. Maps for the Visualization of High – dimensional Data Spaces[C]. In: *Proceedings of Workshop on Self – Organizing Maps*(*WSOM*'2003), Hibikino, Kitakyushu, Japan. 2003:225 – 230.

[13] 刘群等. 基于层叠隐马模型的汉语词法分析[J]. *计算机研究与发展*,2004(8):1421 – 1430.

[14] About SOm Toolbox[EB/OL]. [2008 – 10 – 16] http://www.cis.hut.fi/projects/somtoolbox/about.

[15] SOM Toolbox[EB/OL]. [2008 – 10 – 16] http://www.cis.hut.fi/somtoolbox/package/docs2/som_norm_ variable.html.

[16] Kohonen T. Self – Organizing Maps[M] (3rd ed.). Berlin: Springer, 2001.

[17] Stijn V L, Bert V C, Jeroen M, Bart W, et al. Prediction of Dose Escalation for Rheumatoid Arthritis Patients under Infliximab Treatment [J]. *Engineering Applications of Artificial Intelligence*, 2006, 19(7): 819 – 828.

[18] Kohonen T. Things you haven't heard about the Self – Organizing Map[C]. In: *Proceedings of International Conference on Neural Networks* (*ICNN*' 1993), San Francisco, U. S. A. 1993:1147 – 1156.

[19] Goldensohn S B, Hannan K, McDonald R, et al. Building a Sentiment Summarizer for Local Service Reviews[C]. In: *Proceedings of NLP Challenges in the Information Explosion Era* (*NLPIX*'2008), Beijing, China. 2008:1 – 9.

(作者 E – mail:yuchuanming2003@126.com)