基于 Bayes 潜在语义模型的半监督 Web 挖掘*

宫秀军, 史忠植

(中国科学院 计算技术研究所 智能信息处理开放实验室,北京 100080)
E-mail: {gongxj,shizz}@ics.ict.ac.cn
http://www.ics.ict.ac.cn

摘要: 随着互联网信息的增长,Web 挖掘已经成为数据挖掘研究的热点之一.网页分类是通过学习大量的带有类别标注的训练样本来预测网页的类别,人工标注这些训练样本是相当繁琐的.网页聚类通过一定的相似性度量,将相关网页归并到一类.然而传统的聚类算法对解空间的搜索带有盲目性和缺乏语义特征.提出了两阶段的半监督文本学习策略.第 1 阶段,利用贝叶斯潜在语义模型来标注含有潜在类别主题词变量的网页的类别;第 2 阶段,利用简单贝叶斯模型,在第 1 阶段类别标注的基础上,通过 EM(expectation maximization)算法对不含有潜在类别主题词变量的文档作类别标注.实验结果表明,该算法具有很高的精度和召回率.

关 键 词: 贝叶斯潜在语义分析;半监督学习;简单贝叶斯分类;期望最大化算法;Web 挖掘
中图法分类号: TP393 文献标识码: A

随着互联网的普及,网上信息正在呈指数级增长.合理地组织这些信息,以便从茫茫的数据世界中检索到期望的目标,并有效地分析这些信息,以便挖掘出新颖的、潜在的有用模式,正在成为网上信息处理的研究热点.网上信息的分类目录组织是提高检索效率和检索精度的有效途径,如在利用搜索引擎对网页数据进行检索时,若能提供查询的类别信息,必然会缩小与限制检索范围,从而提高查准率.同时,分类可以提供信息的良好组织结构,便于用户进行浏览和过滤信息.很多大型网站都采用这种组织方式,如 Yahoo[1]采用人工方式来维护网页的目录结构;Google 网站采用一定的排序机制,使与用户最相关的网页排在前面,便于用户浏览.Deerweser[2]等人利用线性代数的知识,通过矩阵的奇异值分解(singular value decomposition,简称 SVD)来进行信息滤波和潜在语义索引(latent semantic index,简称 LSI).它将文档在向量空间模型(VSM)中的高维表示,投影到低维的潜在语义空间(LSS)中,这一方面缩小了问题的规模,另一方面也从一定程度上避免了数据的过分稀疏现象.它在语言建模、视频检索及蛋白质数据库等实际应用中取得了较好的效果.

聚类分析是文本挖掘的主要手段之一[3].它的主要作用是:(1) 通过对检索结果的聚类,将检索到的大量网页以一定的类别提供给用户,使用户能够快速定位期望的目标;(2) 自动生成分类目录;(3) 通过相似网页的归并,便于分析这些网页的共性.K-均值聚类是比较典型的聚类算法,另外,自组织映射(SOM)神经网络聚类和基于概率分布的贝叶斯层次聚类(HBC)等新的聚类算法也正在不断地研制与应用.然而这些聚类算法大部分是一种无监督学习,它对解空间的搜索带有一定的盲目性,因而聚类的结果在一定程度上缺乏语义特征.同时,在高维情况下,选择合适的距离度量标准变得相当困难.而网页分类是一种监督学习,它通过对一系列训练样本的分析来预测未知网页的类别归属.目前已有很多有效的算法来实现网页的分类[4,5],如 Naive Bayesian[6],SVM 等.遗憾的是,获得大量的、带有类别标注的样本的代价是相当昂贵的,而这些方法只有通过大规模的训练才能获得较高精度的分类效果.此外,在实际应用中,分类体系常常是不一致的,这为目录的日常维护带来了一定的困

* 收稿日期: 2001-06-04; 修改日期: 2001-09-06
基金项目: 国家自然科学基金资助项目(60073019,69803010)
作者简介: 宫秀军(1972－),男,内蒙古赤峰人,博士,主要研究领域为数据挖掘,数据仓库技术;史忠植(1941－),男,江苏无锡人,研究员,博士生导师,主要研究领域为数据挖掘,人工智能,机器学习.

难.Kamal Nigam 等人提出从带有类别标注和不带有类别标注的混合文档中分类 Web 网页[7],它只需要部分带有类别标注的训练样本,结合未标注样本含有的知识来学习贝叶斯分类器.

本文的基本思想是:如果知道一批网页 $D=\{d_1,d_2,...,d_n\}$ 是关于某些潜在类别主题变量 $Z=\{z_1,z_2,...,z_k\}$ 的描述,通过引入贝叶斯潜在语义模型,首先将含有潜在类别主题变量的文档分配到相应的类主题中,接着利用简单贝叶斯模型,结合前一阶段的知识,完成对未含类主题变量的文档作标注.针对这两个阶段的特点,我们定义了两种似然函数,并利用 EM(expectation maximization)算法获得最大似然估计的局部最优解.这种处理方法一方面克服了非监督学习中对求解空间搜索的盲目性;另一方面,它不需要对大量训练样本的类别标注,只需提供相应的类主题变量,把网站管理人员从繁琐的训练样本的标注中解脱出来,提高了网页分类的自动性.为了与纯粹的监督与非监督学习相区别,称这种方法为半监督学习算法.

本文第 1 节在简要介绍潜在语义分析(LSA)的基本原理之后,给出了贝叶斯潜在语义模型,并分析了该模型用于 Web 挖掘的优越性.第 2 节主要给出两阶段半监督文本挖掘算法的原理,并针对两种不同的情况分别给出了相应的 EM 算法.第 3 节给出了算法的实验设计和结果的评价分析.最后是文章的结论和进一步研究的展望.

1 Bayes 潜在语义模型

1.1 潜在语义分析

潜在语义分析(latent semantic analysis,简称 LSA)的基本观点是:把高维的向量空间模型(VSM)表示中的文档映射到低维的潜在语义空间中.这个映射是通过对项/文档矩阵 $N_{m\times n}$ 的奇异值分解(SVD)来实现的.具体地说,对任意矩阵 $N_{m\times n}$,由线性代数的知识可知,它可分解为下面的形式:

$$N=U\Sigma V^T. \tag{1}$$

这里,U,V 是正交阵($UU^T=VV^T=I$). $\Sigma=diag(a_1,a_1,...,a_k,...,a_v)$ ($a_1,a_2,...,a_v$ 为 N 的奇异值)是对角阵.潜在语义分析通过取 k 个最大的奇异值,而将剩余的值设为零来近似式(1).

$$\tilde{N}=U\tilde{\Sigma}V^T\approx U\Sigma V^T=N. \tag{2}$$

由于文档之间的相似性,可以通过 $NN^T\approx\tilde{N}\tilde{N}^T=U\tilde{\Sigma}^2U^T$ 来表示,因此文档在潜在语义空间中的坐标可以通过 $U\tilde{\Sigma}$ 来近似.所以,高维空间中的文档表示投影到低维的潜在语义空间中,原来在高维中比较稀疏的向量表示在潜在语义空间中变得不再稀疏.这也暗指,即使两篇文档没有任何共同的项,仍然可能找到它们之间比较有意义的关联值.

通过奇异值分解,将文档在高维向量空间模型中的表示,投影到低维的潜在语义空间中,有效地缩小了问题的规模.潜在语义分析在信息滤波、文本索引、视频检索等方面有较为成功的应用.然而矩阵的 SVD 分解因对数据的变化较为敏感,同时缺乏先验信息的植入等而显得过分机械,从而使它的应用受到了一定的限制.

1.2 Bayes潜在语义模型

经验表明,人们对任何问题的描述都是围绕某一主题展开的.各个主题之间具有相对明显的界限,同时由于偏爱、兴趣等的不同,对不同主题的关注也存在着差别,也就是说,对不同的主题具有一定的先验知识.基于此,我们给出文档产生的潜在贝叶斯语义模型:

设文档集合为 $D=\{d_1,d_2,...,d_n\}$,词汇集为 $W=\{w_1,w_2,...,w_m\}$,则文档 $d\in D$ 的产生模型可表述为

(a) 以一定的概率 $P(d|\theta)$选择文档 d;

(b) 选取一个潜在的类主题 z,该类主题具有一定的先验知识 $p(z|\theta)$;

(c) 类主题 z 含于文档 d 的概率为 $p(z|d,\theta)$;

(d) 在类主题 z 的条件下,产生词 $w\in W$,其概率为 $p(w|z,\theta)$.

经过上述过程获得观测点对(d,w),潜在的类主题 z 被忽略,产生下面的联合概率模型:

$$p(d,w|\theta)=p(d|\theta)p(w|d,\theta),\ p(w|d,\theta)=\sum_{z\in Z}p(w|z,\theta)p(z|d,\theta). \tag{3}$$

该模型是建立在下面的独立性假定基础上的混合概率模型.

(a) 每一观测点对(d,w)的产生是相对独立的,它们通过潜在类主题而相互联系;

(b) 词 w 的产生独立于具体的文档 d,而只依赖于潜在的类主题变量 z.

式(3)也表明,在某一文档 d 中,词 w 的分布是它在潜在类主题下的凸组合,组合权重是该文档类属于此主题的概率.

利用贝叶斯公式和上面的独立性条件,得到

$$p(d,w|\theta)=\sum_{z\in Z}p(z|\theta)p(w|z,\theta)p(d|z,\theta). \tag{4}$$

对应于式(2)的奇异值分解: $U=\{p(d_i|z_k)\}_{n\times k}, V=\{p(w_i|z_k)\}_{m\times k}, \tilde{\Sigma}=diag(p(z_k),p(z_k),...,p(z_k))$,因此 Bayes 潜在语义模型与 SVD 在形式上是统一的,然而与潜在语义分析相比,贝叶斯潜在语义模型具有较为稳固的统计学基础,克服了 LSA 中的数据敏感性;它对潜在的类变量植入先验信息,避免了 SVD 的机械性.

2 半监督文本挖掘算法设计与分析

2.1 算法的一般原理

文本分类是一种监督学习算法,能有效地缩小检索过程中的搜索空间.它通过对大量的带有类别标注的训练样本的学习来预测未知样本的类别归属.目前已经有很多成熟的算法来实现文本分类,获得了较高的精度(precision)和召回率(recall).然而获得分类中带有标注的训练样本的代价是相当昂贵的,因此 Kamal Nigam 等人提出从带有和不带有类别标注的混合文档中学习分类 Web 网页,并获得了很好的精度,但它仍然需要一定量的标注样本.网页聚类通过一定的相似性度量,将相关网页归并到一类,这样也能达到缩小搜索空间的目的,然而传统的聚类方法在处理高维和海量数据时,其效率和精确度却大打折扣.这一方面是由于非监督学习对解空间的搜索本身具有一定的盲目性,另一方面,在高维的情况下很难找到比较适宜的相似性度量标准.例如,欧式距离度量在维数较高时,变得不再适用.基于上面的监督学习与非监督学习的特性,我们提出了一种半监督学习算法.在贝叶斯潜在语义模型的框架下,由用户提供一定数量的潜在类别变量,而不需要任何带有类别标注的样本,将一组文档集划分到不同的类别中.

它的一般模型可以描述为:已知文档集 $D=\{d_1,d_2,...,d_n\}$ 和它的词汇集 $W=\{w_1,w_2,...,w_m\}$,一组带有先验信息 $\theta=\{\theta_1,\theta_2,...,\theta_k\}$ 的潜在类别变量 $Z=\{z_1,z_2,...,z_k\}$,找出 D 上的一个划分 $D_j(j\in(1...k))$,使得 $\bigcup_{j=1}^{k}D_j=D, D_i\cap D_j=\varnothing\ (i\neq j)$.首先,我们将 D 划分为两个集合: $D=D_L\cup D_U$,满足:

$$D_L=\{d\mid \exists j, z_j\in d, j\in[1...k]\}, D_U=\{d\mid \forall j, z_j\notin d, j\in[1...k]\}.$$

我们的算法对文档的类别标注分两个阶段实现:

第 1 阶段.对 D_L 中的元素,我们利用贝叶斯潜在语义模型,在基于 EM 参数估计的基础上,用潜在类别变量来标注文档.即

$$l(d)=z_j=\max_i\{p(d|z_i)\}. \tag{5}$$

第 2 阶段.对 D_U 中的元素,根据对 D_L 中的元素的类别标注,利用 Naïve Bayes 分类模型,经过 EM 算法来实现类别标注.

2.2 对含有潜在类别主题词的文档的类别标注

在理想的情况下,任何文档都不含有两个以上的潜在类别主题词,此时只需将该文档标以潜在的类别.然而在实际应用中,这种理想的情况很难达到.一方面由于很难选择这样的潜在类别主题词.另一方面,这种要求也是不太现实的,因为两个不同的类别很可能含有多个潜在的类别主题词.如在“经济”类的文档中,很可能含有像“政治”、“文化”这样的词.我们的处理方法是把它们分到与潜在的类别主题词语义最为密切的类别中.在我们选择的似然标准下,通过一定次数的 EM 迭代,最后通过式(5)来决定文档的类别.

EM 算法是稀疏数据参数估计的主要方法之一.它交替地执行 E 步和 M 步,以达到使似然函数值增加的目的.在我们的算法中,采用最大化下面的似然函数的方法来实现 EM 迭代.

$$\sum_{d\in D}\sum_{w\in W} n(d,w)\log p(d,w). \tag{6}$$

(a) 在 E 步,用下面的式子估计期望值:

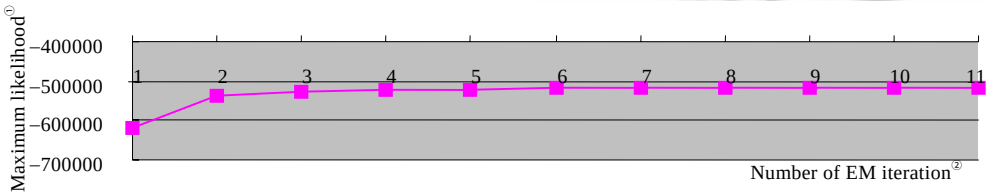
$$P(z\mid d,w)=\frac{p(z)p(d\mid z)p(w\mid z)}{\sum_{z'}p(z')p(d\mid z')p(w\mid z')}. \tag{7}$$

从概率语义上讲,它是用潜在类主题变量 z 来解释词 w 在文档 d 中出现的概率度量.

(b) 在 M 步中,利用上一步的期望值来最大化当前的参数估计.

$$p(w\mid z)=\frac{\sum_d n(d,w)p(z\mid d,w)}{\sum_{d,w'} n(d,w')p(z\mid d,w')},\ p(d\mid z)=\frac{\sum_w n(d,w)p(z\mid d,w)}{\sum_{d',w} n(d',w)p(z\mid d',w)},\ p(z)=\frac{\sum_{d,w} n(d,w)p(z\mid d,w)}{\sum_{d,w} n(d,w)}. \tag{8}$$

相对于潜在语义分析中的 SVD 分解,EM 算法具有线性的收敛速度,并且简单,容易实现,可以使似然函数达到局部最优.图 1 是我们在实验过程中得到的 EM 迭代次数与似然函数值之间的关系.



①似然函数值,②迭代次数.

Fig.1 Relation between maximum likelihood and number of EM iteration

图 1 最大似然估计值与 EM 迭代次数的关系

2.3 基于Naïve Bayes模型学习标注和未标注样本

传统的分类方法都是通过一定的学习机制,在对带有类别标签的训练样本学习的基础上来决定未知样本的类别标签.然而,获得大量的带有类别标注的训练样本,这样的任务是相当繁琐的.Kamal Nigam 等人的研究表明,未带类别标注的文档仍然含有学习分类模型的大量信息.基于此,我们利用简单贝叶斯(Naïve Bayesian)模型作为分类器,把未带标签的训练样本作为一种特殊的缺值状态,通过一定的 EM 迭代算法来估计这种缺值.

在这里,我们首先给出 Naïve Bayes 学习文本分类的一般原理:已知训练文档集 $D=\{d_1,d_2,...,d_n\}$ 和它的词汇集 $W=\{w_1,w_2,...,w_m\}$.每一训练样本是一个 $m+1$ 维向量:$d_i=\langle w_1w_2\ldots w_mc_i\rangle$,其中 $c_i\in C=\{c_1,c_2,...,c_k\}$ 是类别变量.分类的任务就是对未知类别的样本 $d=\langle w_1w_2\ldots w_m\rangle$ 预测它的类别:$c=\max_{j\in[1...k]}\{p(c_j\mid d,\theta)\}$ (这里,θ是模型的参数).

为计算上式,将其展开得到:

$$p(d\mid c_j,\theta)=p(\mid d\mid)\prod_{k=1}^{\mid d\mid}p(w_k\mid c_j;\theta;w_q,q<k). \tag{9}$$

Naïve Bayes 模型在计算式(9)时,引入了下面一些独立性假设:

(a) 文档的词的产生独立于它的内容,即词在文档中出现的位置无先后关系.

(b) 文档中各个词相对于类别属性是相对独立的.

在上面的独立性假定下,结合贝叶斯公式(9)可记为

$$p(d\mid c_j,\theta)=p(\mid d\mid)\prod_{r=1}^{\mid d\mid}p(w_r\mid c_j;\theta;)=\frac{p(c_j\mid\theta)\prod_{r=1}^{m}p(w_r\mid c_j;\theta)}{\sum_{i=1}^{k}p(c_i\mid\theta)\prod_{r=1}^{m}p(w_r\mid c_i;\theta)}. \tag{10}$$

学习的任务变为从数据中利用一定的先验信息来学习模型的参数.在这里,我们选用多项分布模型和 Dirichlet 共轭先验.

$$\theta_{c_j}=\frac{\sum_{i=1}^{\mid D\mid}I(c(d_i)=c_i)}{\mid D\mid},\ \theta_{w_t\mid c_j}=\frac{\alpha_j+\sum_{i=1}^{\mid D\mid}n(d_i,w_t)I(c(d_i)=c_j)}{\alpha_0+\sum_{k=1}^{m}\sum_{i=1}^{\mid D\mid}n(d_i,w_k)I(c(d_i)=c_j)}. \tag{11}$$

这里,$\alpha_0=\sum_{i=1}^{k}\alpha_i$ 为模型的超参数.函数 $c(\cdot)$是类别标注函数,$I(a=b)$为示性函数(若 $a=b$,则 $I(a=b)=1$,否则

$I(a=b)=0$.

尽管 Naïve Bayes 对模型的适用条件作了较为苛刻的限制,然而大量实验表明,即使在违背这些独立性假定的条件下,它仍能表现出相当的健壮性.它已经成为文本类中广为使用的一种方法.

下面我们将通过引入一种最大化后验概率(MAP)似然标准,并结合未标注样本的知识为这些未标注的样本贴标签.

考虑所有的样本集 $D = D_L \cup D_U$,其中 D_L 中的元素在第 1 阶段已被贴上标签.假设 D 中各样本的产生是相互独立的,那么下面的式子成立:

$$p(D \mid \theta) = \prod_{d_i \in D_U} \sum_{j=1}^{|C|} p(c_j \mid \theta) p(d_i \mid c_j; \theta) \cdot \prod_{d_i \in D_L} p(c(d_i) \mid \theta) p(d_i \mid c(d_i); \theta) . \tag{12}$$

在上面的式子中,将未标注的文档看做是混合模型.我们的学习任务仍然是通过样本集 D 来获得模型参数 θ 的最大估计.利用贝叶斯定理,取对数后验似然得:

$$l(\theta \mid D) = \log p(\theta \mid D) = \log \frac{p(\theta)}{p(D)} + \sum_{d_i \in D_U} \log \sum_{j=1}^{|C|} p(c_j \mid \theta) p(d_i \mid c_j; \theta) + \sum_{d_i \in D_L} \log p(c(d_i) \mid \theta) p(d_i \mid c(d_i); \theta) . \tag{13}$$

为估计未标注样本的标签,借用前一节潜在语义分析中的潜在类主题变量的思想,我们在这里引入 k 个潜在变量 $Z = \{z_1, z_2, ..., z_k\}$,每个潜在变量是 n 维向量 $z_i = \langle z_{i1}, z_{i2}, \ldots, z_{in} \rangle$,并且,如果 $c(d_j) = c_i$,那么 z_{ij}=1,否则 z_{ij}=0.这里的潜在变量与前面提到的潜在类主题变量表示的意义是不同的.这里的潜在变量是为了表示的方便而引入的,它独立于文档的特征;而潜在类主题变量则隐藏在文档的特征之中,是可以根据经验指定的.式(13)可以统一表示成以下形式:

$$l(\theta \mid D) = \log \frac{p(\theta)}{p(D)} + \sum_{i=1}^{|D|} \sum_{j=1}^{|C|} z_{ji} \log p(c_j \mid \theta) p(d_i \mid c_j; \theta_j) . \tag{14}$$

在式(14)中,已标注的样本 z_{ji} 是已知的,学习的任务是最大化模型的参数和对未知的 z_{ji} 的估计.在这里,我们采用式(14)为似然函数,仍然用 EM 算法来学习未标注样本的知识.但它的过程与前一阶段有所不同.

在 E 步的第 k 次迭代中,基于当前的参数估计,利用简单贝叶斯分类器来计算未标注样本的类别.对

$$\forall d \in D_u : p(d \mid c_j, \theta^k) = \frac{p(c_j \mid \theta^K) \prod_{r=1}^{m} p(w_r \mid c_j; \theta^k)}{\sum_{i=1}^{k} p(c_i \mid \theta^k) \prod_{r=1}^{m} p(w_r \mid c_i; \theta^k)}, \ j \in [1...k] ,$$

取获得最大后验概率的类别 c_i 作为该文档的期望类别标注,即 $z_{id} = 1$, $z_{jd} = 0$ $(j \neq i)$.

在 M 步,基于前一步获得的期望值,最大化当前的参数估计:

$$\theta_{c_j} = p(c_j \mid \theta) = \frac{\sum_{i=1}^{|D|} z_{ji}}{|D|}, \theta_{w_t \mid c_j} = p(w_t \mid c_j; \theta) = \frac{\alpha_j + \sum_{i=1}^{|D|} n(d_i, w_t) z_{ji}}{\alpha_0 + \sum_{k=1}^{m} \sum_{i=1}^{|D|} n(d_i, w_k) z_{ji}} . \tag{15}$$

3 实验设计与结果分析

我们的实验数据是用 Spider 从 http://www.fm365.com 搜集的关于体育方面的网页,在每一类别中都包括了含有类别词的网页和不含有类别词的网页.它们的类别和分布见表 1.

Table 1 Taining examples and their distributions

表 **1** 选择的训练文档及其分布

	Football	Basketball	Volleyball	Table tennis	Tennis	Chess and cards
Has subject words①	40	60	48	30	57	80
No subject words②	80	40	29	11	6	4

①含有主题词,②不含有主题词.

实验中共有 485 篇网页,经过切词处理,去掉一定的停用词后,共计有 2 719 个词,经过半监督学习算法处理,得到下面的结果,见表 2.

Table 2 Results of semi-supervised Web mining

表 2 半监督 Web 挖掘的结果

	Football	Baseketball	Volleyball	Table tennis	Tennis	Chess and cards	Result evaluate①	
							Precision②	Recall③
Football(120)	112	5	0	1	2	2	0.965 52	0.933 33
Basketball(100)	1	98	0	1	0	0	0.933 33	0.980 00
Volleyball(77)	1	2	74	0	0	0	0.973 68	0.961 03
Table tennis(41)	0	0	0	40	1	0	0.833 33	0.975 61
Tennis(63)	0	0	1	4	58	0	0.935 48	0.920 63
Chess and cards(84)	4	0	1	2	1	76	0.974 36	0.904 76

①结果评价,②精度,③召回率.

在足球类的 120 篇文档中,分到篮球、排球、乒乓球、网球及棋牌类的文档个数分别为 5,0,1,2,2.其中分到篮球类的文档数较多,通过进一步的研究我们发现,这是由于这些文档中含有的词与篮球类中含有的词中同义词较多所致.我们使用同样的数据,利用 Naïve Bayes 分类,得到类似的结果.

另外,选取足球类的文档 1 000 篇,经过初步的预处理后得到 876 个词.图 2 是选取不同的潜在类别变量分到各类中的文档个数对比.

第 1 次选择 14 个潜在变量,第 2 次选择 7 个潜在变量,两次的差别是,在第 1 次选择中“甲 A”的各个俱乐部在第 2 次中用“甲 A”来替代.从图中可以看出,第 1 次选择各个类别分到的文档数基本相同,而在第 2 次选择中,分到“甲 A”中的文档数近似为各“俱乐部”中的文档树之和,分到其余类别中的文档数基本不变.这个结果与我们的抽样基本上是吻合的,同时也说明,潜在类别变量的概括能力具有一定的层次性.

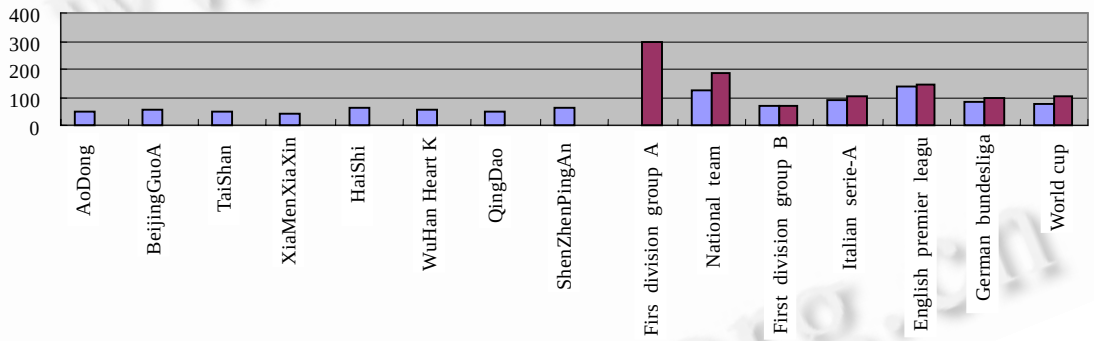


Fig.2 Different selection of latent variable and document's distribution

图 2 不同潜在变量的选择及各类别的文档分配

4 结论与展望

网上信息的分类目录组织是提高检索效率和检索精度的有效途径.它通过学习大量的带有类别标注的训练样本来预测网页的类别,然而人工标注这些训练样本是相当繁琐的.网页聚类通过一定的相似性度量,将相关网页归并到一类,也能达到缩小搜索空间的目的,然而传统的聚类方法对解空间的搜索带有盲目性和缺乏语义特性,因而它的效率和精确度大打折扣.我们提出了一种半监督学习算法.在贝叶斯潜在语义模型的框架下,由用户提供一定数量的潜在类别变量,而不需要任何带有类别标注的样本,将一组文档集划分到不同的类别中.它分为两个阶段:第 1 阶段,利用贝叶斯潜在语义分析来标注含有潜在类别变量的文档的类别;第 2 阶段则通过简单贝叶斯模型,结合未标注文档的知识,对这些文档贴标签.我们还分析了算法的时间与空间复杂性,实验结果也表明,该算法具有较高的精度与召回率.对该算法的进一步研究包括潜在类别变量的选择对结果的影响以及在贝叶斯潜在语义分析框架下如何实现词的聚类等.

References:

[1] Lan, Huang. A survey on web information retrieval technologies. http://citeseer.nj.nec.com/cache/papers2/cs/16461/http:zSzzSzwww.ecsl.cs.sunysb.eduzSztrzSzrpe8.pdf/a-survey-on-web.pdf.

[2] Deerwester, S., Dumais, S.T., G.W., *et al*. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990,41.

[3] Shi, Zhong-zhi. Knowledge Discovery. Beijing: Tsinghua University Press, 2000. (in Chinese).

[4] Li, Xiao-li, Liu, Ji-min, Shi, zhong-zhi. Concept inference network and its application in text classification. Computer Research and Development, 2000,37(9):1032~1038 (in Chinese).

[5] Shivakumar, Vaithyanathan. Hierarchical Bayes for text classification. In: Tan, Ah-Hwee, Yu, P.S., eds. Proceedings of the International Workshop on Text and Web Mining. 2000.

[6] Chickering, D., Heckerman, D. Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. Technical Report, MSR-TR-96-08, Microsoft Research, 1996.

[7] Nigam, K., McCallum, A., Thrun, S., *et al*. Learning to classify text from labeled and unlabeled documents. In: Mostow, J., Madison, C.R., eds. Proceedings of the 15th National Conference on Artificial Intelligence. Wisconsin: AAAI Press, 1998. 792-799.,

附中文参考文献:

[3] 史忠植.知识发现.北京:清华大学出版社,2000.

[4] 李晓黎,刘继敏,史忠植.概念推理网及其在文本分类中的应用.计算机研究与发展,2000,37(9):1032~1038.

Semi-Supervised Web Mining Based on Bayes Latent Semantic Model*

GONG Xiu-jun, SHI Zhong-zhi

(*Laboratory of Intelligent Information Processing, Institute of Computing Technology, The Chinese Academy of Sciences, Beijing* 100080, *China*)

E-mail: {gongxj,shizz}@ics.ict.ac.cn

http://www.ics.ict.ac.cn

Abstract: With the increasing of information on Internet, Web mining has been the focus of data mining. Web classification predicts the labels of Web documents by learning lots of training examples with labels. It is very expensive to get these examples by manual. Web clustering groups the similar Web documents by a certain of metric of similarity. But the classical algorithms of clustering are aimless in searching the solution space and absent of semantic characters. In this paper, a semi-supervised learning strategy consists of two stages is put forward. The first stage, labels the documents that include latent class variables by using Bayes latent semantic model. The second stage, based on the results from the first stage, labels the documents excluding latent class variables with the Naïve Bayes models. Experimental results show that this algorithm has good precision and recall rate.

Key words: Bayes latent semantic analysis; semi-supervised learning; Naïve Bayesian classifier; expectation maximization; Web mining

* Received June 4, 2001; accepted September 6, 2001
Supported by the National Natural Science Foundation of China under Grant Nos.60073019, 69803010