基于序列的文本自动分类算法*

解冲锋, 李 星

(清华大学 电子工程系,北京 100084)
E-mail: xcf@public.bjnet.edu.cn; xing@cernet.edu.cn
http://www.tsinghua.edu.cn

摘要: 提出了一种基于序列的文本自动分类算法.该算法利用了文本中两个层次的语义相关性:句子(子模式)之间的相关性和句子内代表特定含义的关键词(概念节点)之间的相关性,这样就实现了对关键词的动态加权.对于不含有关键词的子模式,采用 Markov 模型来对其信号幅度进行估计,从而生成一个待分类文本的特征序列.在中文文本分类实验中,可以达到 83%的 BEP 值.此外,该算法在实际系统中容易实现.

关 键 词: 序列;概念节点;自动分类;相关度
中图法分类号: TP18 文献标识码: A

文本自动分类就是对大量的用自然语言写成的文本按照一定的主题类别自动进行分类.文本分类是信息处理的一个重要分支,在信息发现领域中有着重要的用途,特别是在网络技术飞速发展的时代,对网络上的海量网页文本进行过滤和分类可使用户快速发现真正有用的文本.文本分类算法很多,典型的有基于实例[1]、Sleeping expert[2]、基于推理网络[3]以及基于规则组[4]等算法.这些算法一般需通过大量的训练才能获得较好的效果,如在基于实例的分类算法中,为了获得主题类别和文本之间的相关度,需要用大量的样本来获得关键词的权值,这样的算法在实际系统中实现代价较大.本文提出了一种基于序列的文本自动分类算法(简称序列算法),这个算法利用了文本内两个层次的语义相关性:句子之间的相关性和句子内代表特定含义的关键词之间的相关性,从而实现了对关键词的动态加权.在对汉语文本进行分类的实验中,它可以达到较好的分类正确率,而且与其他分类算法相比,本算法在实际系统中容易实现.

1 定 义

在本算法中,称一个待分类的文本为未知文本.设有 M 个未知文本,其中第 i 个未知文本为 T_i.在 T_i 中包含 N 个子模式,第 j 个子模式 s_j 可以是 T_i 中任意完整的题目、标题或句子,其中序号 j 表示子模式在 T_i 中的位置.在子模式 s_j 中定义概念节点 p_k,它是子模式内关键词 ws_k 当前代表的含义,即 $Meanings(ws_k|s_j)=p_k$.其次,C_l 表示第 l 个主题类别,本地字典 D_l 是含有主题类别 C_l 的各种关键词的字典[4],其中的每个关键词 w_n 代表了它在本类 C_l 内的含义,即 $D_l=\{w_n|Category(Meanings(w_n))=C_l\}$.本地字典是判断 T_i 与 C_l 相关度的原始知识,其生成属于类别特征提取过程,典型的特征提取方法有基于互信息的特征提取算法、Sleeping expert 算法、从专业词典中获取以及在训练语料不足的情况下可直接采用人工添加的方法.未知文本、子模式和概念节点的包含关系如图 1 所示.

* 收稿日期: 2000-08-01; 修改日期: 2000-10-30
基金项目: 国家"九五"重点科技攻关项目(96-743-01-05-01)
作者简介: 解冲锋(1974－),男,陕西咸阳人,博士生,主要研究领域为信息检索,计算机网络;李星(1956－),男,北京人,博士,教授,博士生导师,主要研究领域为信息检索,信号处理,计算机网络数据库理论.

2 词义互相激励原则

文本分类的首要问题是:当子模式内关键词 ws_k 和 D_l 内关键词 w_n 相匹配时,即 $w_n=ws_k$,它们的含义具有多大的相似性,即 $Similarity(Meanings(ws_k),Meanings(w_n))$如何度量的问题.这个问题的产生是由于自然语言中许多词汇具有多义性,即一个词汇可以表示多个概念,如图 2 所示,w_k 表示一个多义词,则满足 $Meanings(w_k)=c_r$ 的 c_r 有多个,分别为 $c_1,c_2,\ldots,c_n$.例如关键词"病毒"既是一个计算机词汇,又是一个生物学词汇.多义词的存在使得未知文本和本地字典中词形匹配的两个词可能具有不同的含义,即代表不同的概念.对于这种情况,传统的作法通常用权重来解决这个问题[1~3],但这种权重是静态的,并且需要大量的训练来获得每个关键词的权重值,而很少考虑上下文的作用.实际上,上下文在判定词汇含义的过程中具有重要作用,例如,在一个包含"计算机"、"软件"、"磁盘"等词汇的子模式中,"病毒"代表生物学上病毒含义的概率很小,而在包含"基因"、"生物"、"细菌"等词汇的子模式中则相反.它说明了聚集在一个小单元(例如子模式)内的可代表某类别含义的词越多,每个词代表本类含义的可能性就越大,从形式上表现为同类关键词之间的互相激励,这种互相激励现象在自然语言中是非常普遍的,我们称之为词义互相激励原则.

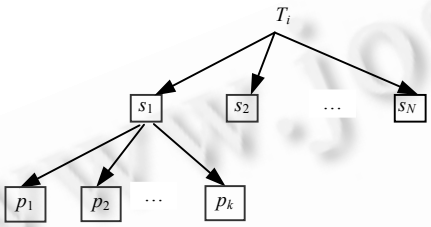


Fig.1 The inclusion relation among unknown text, subpatterns and concept nodes

图 1 未知文本、子模式和概念节点的包含关系

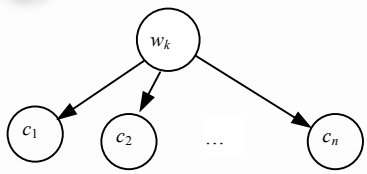


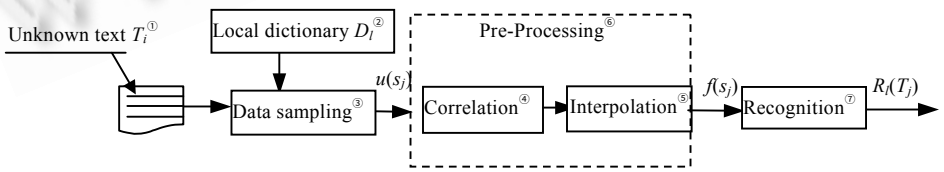
Fig.2 The illustration of polysemy

图 2 关键词多义性的图解

实际上,子模式内不同的词汇的语法作用是不同的,因而在全句的语义作用也是不同的,但在本文中为了突出序列在解决文本分类中的作用,在处理过程中把它们的作用同等看待,这样可以简化概念和算法;其次假定所有的单个词汇在判定未知文本的类别时都是不可靠的,因此采用动态权重来描述它们的含义.序列算法以本地字典作为基础知识,利用同类关键词在子模式内的互相激励来定义本子模式的幅值,称之为子模式对特定类别的激励值,子模式的激励值代表了它与类别在语义上的相关程度.其次,我们也认识到,子模式之间在语义上存在着很强的相关性,可以利用这种相关性估计临近的子模式与类别的语义相关度.

3 分类算法

对文本进行分类的核心是判断未知文本与每一个主题类别在内容上的相关程度[2,5],基于序列的分类算法首先获得每个子模式对特定类别的独立激励值,形成初始序列.然后,根据不同子模式之间的语义相关性对序列中所有子模式的独立激励值进行预处理,这样得到的序列信号可以作为判断未知文本与特定类别相关度的特征表示.因此,把对未知文本的相关度的计算分为以下几个阶段,如图 3 所示.



①未知文本 T_i,②本地字典 D_l,③数据采样,④相关处理,⑤插值平滑,⑥预处理,⑦识别.

Fig.3 The processing of text classification based on sequence

图 3 基于序列的文本分类过程

(1) 数据采样

数据采样主要是获得每个子模式对特定类别的激励值.该过程首先把本地字典 D_l 中每个本类关键词 w_n 和子模式 s_j 进行比较,在获得匹配的关键词对集合 $G_j=\{(w_n,ws_k)|w_n=ws_k,w_n\in D_l,ws_k\in s_j\}$ 后,再对子模式中每个关键词对应的概念节点 $p_k=Meanings(ws_k|s_j)$ 进行分析,从而计算出子模式 s_j 的激励值.子模式的激励值与以下因素有关:首先是集合 G_j 内元素的数目 $N(s_j)=sizeof(G_j)$;其次是每个 $Similarity(Meanings(w_n),p_k)$,即 s_j 中每个匹配关键词的概念节点和它的匹配对象在语义上的相似度;最后是每个匹配关键词的词频 $f(ws_k).N(s_j)$ 和 $f(ws_k)$ 可以直接获得,s_j 内关键词和它的匹配对象在语义上的相似度 $Similarity(Meanings(w_n),p_k)$ 可通过互相激励原则来表示,这就是说,G_j 内的元素数 $N(s_j)$ 越大,则 $Similarity(Meanings(w_n),p_k)$ 越大,若用线性关系来表示,则有

$$Simarility(Meanings(w_n),p_k)=\sigma N(s_j). \tag{1}$$

从式(1)可以看出,每个匹配对的相似度随其所在子模式的不同会做相应变化.综合以上 3 个因素,则子模式的激励为

$$\begin{aligned} u(s_j)&=\sum_{k=1}^{N(s_j)} Simarility(Meanings(w_n),p_k)f(ws_k)\\ &=\sum_{k=1}^{N(s_j)}\sigma N(s_j)f(ws_k)\\ &=\sigma N(s_j)\sum_{k=1}^{N(s_j)}f(ws_k). \end{aligned} \tag{2}$$

式(2)说明,$N(s_j)$ 越大,则每个匹配关键词的权重即得到提升,反之则相反.在得到每个子模式的激励值后,得到一个如图 4(a)所示的序列.根据序列的幅度的不同,子模式分为两类:强子模式($u(s_j)>0$)和弱子模式($u(s_j)=0$).以下文为例,利用计算机类本地字典进行采样,表 1 中列出了子模式中匹配的关键词的词频和动态权重以及每个子模式的激励值.

“香港即将首次推出互联网保险业务,以减低网络公司面对的各种风险./黑客现象令电子商务的保安问题备受关注,因此使得除电脑系统保安公司正设计各类保安软件外,保险公司也把眼光投向了这个市场./推出此项业务的美亚保险公司的服务对象包括:互联网上做广告的公司、电子商务公司、互联网服务供应商或入门网站及电子科技服务等./这家公司提供的保障范围包括:系统保安不足导致的资料外泄,电脑病毒入侵;因电子商务中断而引致的财产及收入损失;电讯产品服务未能达到应有服务程度导致无形损失等./公司在承保前会先向投保公司进行电脑保安和软硬件系统的测试,目前公司暂时最高保额为 2 500 万美元./美亚保险东南亚区财务保险部副总裁李振威表示,在全球35个国家中,有64%的科技企业对自身的电脑系统并无信心,去年入侵电脑个案大幅上升,相信企业对这项保险业务有一定需求.”

Table 1 The result of data sampling

表 **1** 数据采样的结果

子模式编号 j	关键词(频率,权重)	子模式激励 $u(s_j)$
1	互联网(1,2),网络(1,2)	4
2	黑客(1,4),电子商务(1,4),电脑(1,4),软件(1,4)	16
3	互联网(2,3),电子商务(1,3),网站(1,3)	12
4	系统(1,4),电脑(1,4),病毒(1,4),电子商务(1,4)	16
5	电脑(1,2),硬件(1,2)	4
6	电脑(2,1)	2

(2) 相关处理

强子模式之间存在着语义联系,所以在未知文本内部,一个子模式的较大的激励值同时也是另外一个子模式与类别相关的证据,因此有必要进行强子模式之间相关性的处理,方法如下:

$$f(s_j)=u(s_j)+\sum_{k=1}^{N}u(s_k)\exp(-\mu_t|k-j|),\quad k\neq j. \tag{3}$$

μ_t 为取值为常数的衰减系数.从式(3)可以看出,在文本中相距较近的强子模式之间相互影响较大,这是因为相距较近的子模式相关性较强,多个强子模式与类别有着较强的相关性,从而使得每一个子模式与该类别相关的可信度增强,表现在输出信号中是幅度增加,如图 4(b)所示.

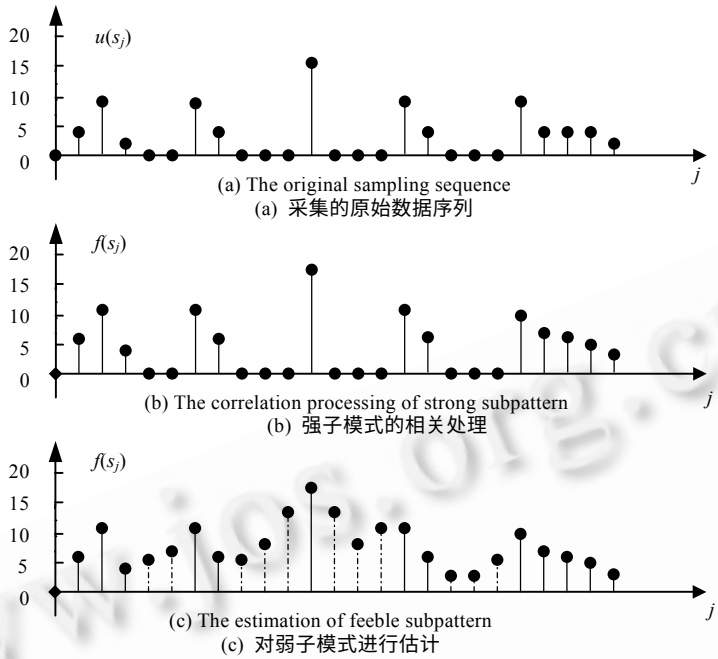


Fig.4 The processing of the sampling data

图 4 采集数据的处理过程

(3) 插值平滑

为了获得弱子模式的信号幅度值,需要根据强子模式的信号幅度对弱子模式进行插值估计.弱子模式中没有本类概念节点并不表示弱子模式与类别绝不相关,从语义上来说它与类别的相关度是由强子模式,即由它所在的上下文来决定的,因此可用插值的方法根据上下文来估计弱子模式的信号强度.弱子模式按照所在环境的不同又分为 3 种:

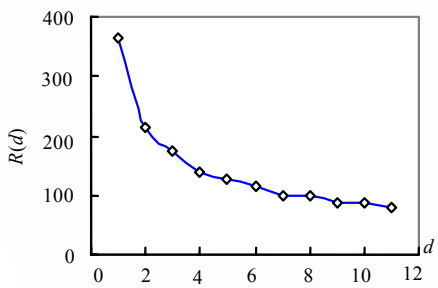


Fig.5 The correlation function of the sequence after phase 2

图 5 第 2 步处理后序列的相关函数

(1) 前弱子模式.位于第 1 个强子模式之前的弱子模式.

(2) 中弱子模式.位于两个强子模式之间的弱子模式.

(3) 后弱子模式.位于最后一个强子模式之后的弱子模式.

对不同弱子模式的信号幅度采用不同的幅度估计算法,中弱子模式的估计算法如下:

经过实验发现,在一个文本序列内部,强子模式之间的幅度相关性与距离有关,如图 5 所示,且相关函数的曲线形状近似为负指数,因此可以把文本序列看做为一个均值非 0 的马尔可夫过程[6],即强子模式之间的相关函数为

$$R_s(k) = \sigma^2 \exp(-\mu_s \mid k \mid) + A^2, \tag{4}$$

其中 A 为强子模式幅度的均值, σ^2 为方差,这两个参数可以通过统计获得.设弱子模式的方差和均值与强子模式相同,则根据邻接强子模式 s_r 和 s_p 估计弱子模式 s_j 的公式如下:

$$\hat{f}(s_j) = \alpha f(s_r) + \beta f(s_p), \quad r < j < p, \tag{5}$$

其中 s_r 是位于 s_j 前面的邻接强子模式,s_p 是位于 s_j 后面的邻接强子模式:

$$\alpha = \frac{(\sigma^2 + A^2)[\sigma^2 \exp(-\mu_s|r-j|) + A^2] - [\sigma^2 \exp(-\mu_s|p-r|) + A^2][\sigma^2 \exp(-\mu_s|p-j|) + A^2]}{(\sigma^2 + A^2)^2 - [\sigma^2 \exp(-\mu_s|p-r|) + A^2]^2}, \tag{6}$$

$$\beta = \frac{(\sigma^2 + A^2)[\sigma^2 \exp(-\mu_s|p-j|) + A^2] - [\sigma^2 \exp(-\mu_s|p-r|) + A^2][\sigma^2 \exp(-\mu_s|r-j|) + A^2]}{(\sigma^2 + A^2)^2 - [\sigma^2 \exp(-\mu_s|p-r|) + A^2]^2}. \quad (7)$$

证明:由上面的前提可知,中弱子模式的幅度由它两边的强子模式进行估计,从而对序列进行平滑.在此采用线性最小均方估计,即规定

$$\hat{f}(s_j) = \alpha f(s_r) + \beta f(s_p). \quad (8)$$

要求选择适当的系数 α 和 β,使误差 $E\{[f(s_j) - \alpha f(s_r) - \beta f(s_p)]^2\}$最小.其中 $f(s_j)$为弱子模式 s_j 的真正幅度.

现在分别求 α 和 β 的值.根据正交性原理,上述 α 和 β 必须使误差与数据 $f(s_r)$和 $f(s_p)$正交,即

$$E\{[f(s_j) - \alpha f(s_r) - \beta f(s_p)] \times f(s_r)\} = 0, \quad (9)$$

$$E\{[f(s_j) - \alpha f(s_r) - \beta f(s_p)] \times f(s_p)\} = 0. \quad (10)$$

由式(9)得

$$R_s(j,r) = \alpha R_s(r,r) + \beta R_s(p,r). \quad (11)$$

由于已经假定序列信号为平稳过程,所以式(11)可以写成:

$$R_s(j-r) = \alpha R_s(0) + \beta R_s(p-r). \quad (12)$$

同理,可由式(10)得

$$R_s(j-p) = \alpha R_s(r-p) + \beta R_s(0). \quad (13)$$

结合式(11)和式(12),得到

$$\alpha = \frac{R_s(0)R_s(r-j) - R_s(r-p)R_s(p-j)}{R_s^2(0) - R_s^2(r-p)}, \quad (14)$$

$$\beta = \frac{R_s(0)R_s(p-j) - R_s(r-p)R_s(r-j)}{R_s^2(0) - R_s^2(r-p)}. \quad (15)$$

将式(4)代入式(14)和式(15)可以得到 α 和 β 的取值,如式(6)和式(7)所示. □

对前弱子模式,同样采用线性最小均方估计,与上面估计方法不同的是,这里是根据最近的一个强子模式进行估计,即

$$\hat{f}(s_j) = f(s_p)\frac{A^2 + \sigma^2 \exp(-\mu_s | j-p |)}{A^2 + \sigma^2}, \qquad j < p. \quad (16)$$

其中 s_p 是距离子模式 s_j 最近的强子模式.上式的证明思路与式(4)相似,限于文章篇幅,这里不再证明.对于后子模式,可以采用和前子模式类似的方法进行处理,这里不再重复.

得到弱子模式幅度的估计值 $f(s_j)$以后,用其代表弱子模式的真正幅度,即 $f(s_j)=f(s_j)$.

(4) 识　别

经过以上的处理,得到了一个如图 4(c)所示的串行序列,如果把文本作为一个未知系统,而把本地字典作为测试信号,那么得到的串行序列可以作为该未知文本的输出信号,这个序列信号可以作为文本的一个重要特征,它的强度反映了在本类别下未知文本内不同子模式所产生的信号强弱,可以作为文本分类的一个判别依据.首先根据序列长度求得未知文本信号的平均幅度,如下

$$A_l(T_i) = \sum_{j=1}^{N} f(s_j) \times \frac{1}{N}. \quad (17)$$

为了能明确地体现相关度,把未知文本的幅度平均值 A_i 进行归一化,从而得到未知文本 T_i 和类别 C_l 之间的相关度:

$$R_l(T_i) = \mathrm{ctg}(A_l(T_i) \times \mu_l) \times 2/\pi. \quad (18)$$

式(18)中 μ_l 为变换系数,其取值为常数.选取门限参数 $\theta_c \in (0,1)$,如果 $R_l(T_i) > \theta_c$,则判定为 l 类,否则为

非 l 类.

4 实验结果

为了验证序列算法的分类性能,在此把它和 Sleeping expert 算法的分类结果进行比较.衡量文本分类效果的指标是查全率(recall)和查准率(precision),其中查全率是被判定为相关的相关文本占全部相关文本的比率;查准率是被判定为相关的文本中真正相关的文本所占的比率.当查全率和查准率相等时,它们的值称为 BEP(breakeven point).

本实验采用的测试语料是《人民日报》网络版 2000 年 1 月~6 月的所有新闻的集合,它共包含 14 440 篇标注好类别的文本,称之为 RMRB-14400.由于部分类别样本数量太少而不具有测试的可靠性,因此只选取教育、法律、音乐和医疗保健 4 个有代表性的类别进行测试,测试语料中属于这 4 个类别的文本数分别为 1 028、1 022、258 和 1 009.每个类别的本地字典是从相应的专业字典中获得,其中含有的元素个数分别为 137、114、165 和 150.测试采用的参数值为 μ_l=2,σ=1,μ_p=2,μ_r=1.2.在多个测试类别的情况下,需采用微平均[7]的方法来获得 Recall 和 Precision 的总体测试结果.

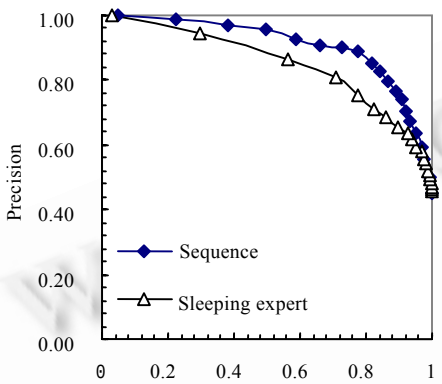


Fig.6 The comparison of performance of two algorithms

图 6 两种算法的性能比较

由于 Sleeping expert 算法需训练过程来获得每个关键词的权重,我们把 RMRB-14400 作为它的训练语料和测试语料.在用两种算法分别获得所有样本与每个类别的相关度后,采取门限均匀变化的方法得到各自的分类结果,图 6 为两种算法得到的 Precision-Recall 曲线.从中可以看出,序列算法的 BEP 值可以达到 83%,明显优于 Sleeping expert 算法;其次,在查全率和查准率接近相等的范围内,在序列算法得到的曲线上数据点分布比较稠密,说明了在这个范围内即使判决门限有较大的变化,查全率和查准率的变化却不显著,这为选择合理的门限提供了方便,而 Sleeping expert 算法不具有这个特性.

5 结束语

Chidandand 利用本地字典来提取规则[4],然后用规则对文本的类别进行识别,他得出的重要结论是,本地字典方法比全局字典方法具有更多的优点,因此序列算法也采用了本地字典作为类别的特征表示.序列算法省掉了训练过程,但在分类的过程中充分利用文本的结构信息,并获得了较好的分类效果.总之,基于序列的分类算法以未知文本的各种标题和句子为子模式,以本地字典为已有知识,根据子模式之间和子模式内概念节点之间的语义联系,生成一个串行序列,该串行序列可以作为对未知文本进行类别判定时的特征表示.本算法的最大特点是,对子模式中概念节点和本地字典中匹配的关键词采取了动态加权的方法,使权重和关键词含义尽量匹配.实验表明,它具有较好的分类性能,而且在实际系统中容易实现.

References:

[1] Chute, C.G. An example based mapping method for text categorization and retrieval. ACM Transactions on Information System, 1994,12(3):252~277.

[2] Cohen, W.W., Singer, Y. Context-Sensitive learning methods for text categorization. ACM Transactions on Information System, 1999,17(2):141~173.

[3] Turle, H., Croft, B. Evaluation of an inference network net-based retrieval model. ACM Transactions on Information System, 1991,9(3):187~222.

[4] Apte, C., Damerau, F. Automated learning of decision rules for text categorization. ACM Transactions on Information System, 1994,12(3):233~251.

[5] Belkin, N.J., Croft, W.B. Information filtering and information retrieval: two sides of the same coin? Communications of the ACM, 1994,35(12):29~38.

[6] Xiang, Jing-cheng, Wang Yi-qing. Singal Detection and Estimation. Beijing: Electronics Industry Press, 1994. 165~166 (in Chinese).

[7] Lam, W., Ruiz, M., Srinivasan, P. Automatic text categorization and its application to text retrieval. IEEE Transactions on Knowledge and Data Engineering, 1999,11(6):865~879.

附中文参考文献:

[6] 向敬成,王意清.信号检测与估计.北京:电子工业出版社,1994.165~166.

A Sequence-Based Automatic Text Classification Algorithm*

XIE Chong-feng, LI Xing

(*Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)

E-mail: xcf@public.bjnet.edu.cn; xing@cernet.edu.cn

http://www.tsinghua.edu.cn

Abstract: An automatic text-classification algorithm based on sequence is presented in this paper. It utilizes the semantic relevance on two levels: relevance between sentences (subpattern) and between keywords which represent specific meaning (concept node) in one sentence. In this way, each keyword can be combined with dynamic weight. For subpatterns which contain no keywords, Markov model is used to estimate the amplitude of their signals, thereby the feature sequence for the text which needs to be classified is created. In the experiment of classifying Chinese documents, it is BEP value is about 83%. Furthermore, it is easy to implement in actual system.

Key words: sequence; concept node; automatic classification; relevant degree

* Received August 1, 2000; accepted October 30, 2000
Supported by the Key Sci-Tech Project of the National 'Ninth Five-Year-Plan' of China under Grant No.96-743-01-05-01