

# 基于特征词复合权重的关联网页分类

兰 均<sup>1</sup> 施化吉<sup>1</sup> 李星毅<sup>1</sup> 徐 敏<sup>2</sup>

(江苏大学计算机科学与通信工程学院 镇江 212013)<sup>1</sup> (南通大学计算机科学与技术学院 南通 226019)<sup>2</sup>

**摘 要** 针对网页分类中关联分类方法存在的如下两点不足:(1)仅把网页当成纯文本处理,忽略了网页的标签信息,(2)仅用网页中的特征词作为关联规则的项,没有考虑特征词的权重,或仅以词频来量化权重,忽略了特征词位置特征的影响,提出了基于特征词复合权重的关联网页分类方法。该方法利用网页标签信息所体现的位置特征计算特征词的复合权重,并以此权重为基础建立分类规则,对网页进行分类。实验结果表明,该方法取得了比传统的关联分类方法更好的效果。

**关键词** 网页分类,关联规则,位置特征,复合权重

中图法分类号 TP391 文献标识码 A

## Associative Web Document Classification Based on Word Mixed Weight

LAN Jun<sup>1</sup> SHI Hua-ji<sup>1</sup> LI Xing-yi<sup>1</sup> XU Min<sup>2</sup>

(School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China)<sup>1</sup>

(School of Computer Science and Technology, Nantong University, Nantong 226019, China)<sup>2</sup>

**Abstract** There are two shortages when the method of classification based on association rules is applied to classify the Web documents; one is that the method process the Web document as a plain text, ignoring the HTML tags information of the Web page; another is that either item of the association rules is only the words in the Web page, without considering the weight of the words, or it quantifies the weight of the word frequency, ignoring the importance of the location of the words in the Web document. Therefore, a new efficient method was proposed in the paper. It calculates the word's mixed weight by the information of the HTML tags feature, and then mines the classification rules based on the mixed weight to classify the Web pages. The result of experiment shows that the performance of this approach is better than the traditional associated classification methods.

**Keywords** Web document classification, Association rules, Location feature, Mixed weight

## 1 引言

随着网络的快速发展,因特网上的网页数量激增,所蕴含的信息量急剧膨胀。利用搜索引擎进行检索时,结果信息庞大且无关信息众多,难以满足用户需求。如何更好地搜索、过滤网络提供的海量资源,成为一个亟待解决的问题。为此,网页分类技术<sup>[1]</sup>应运而生,并逐渐成为继文本分类之后机器学习领域的研究热点。

目前,网页分类技术的研究主要以文本分类的方法为基础,如支持向量机(SVM)<sup>[2]</sup>、贝叶斯分类方法(NB)<sup>[3]</sup>、K邻近算法(KNN)<sup>[4]</sup>等。这些方法取得了一定效果,但仍没达到用户的期望。关联分类方法是近年来出现的一种新的分类方法,它具有分类预测准确度高特点。一些文献对其进行了研究。如文献[5,6]的方法,其基本思想是利用现有关联规则挖掘算法产生所有频繁项目集,并使用这些频繁项目集构造分类器,但其仅是针对非文本分类领域设计的。文献[7,8]将

文本视为事务,特征词作为项,挖掘分类规则,对文本进行分类,获得了较好的准确率。但其用以构造关联规则的项仅是文本中的特征词,忽略了特征词的权重大小。文献[9]虽引入了词频作为特征词的权重,提出了加权频繁项集的概念和相应的加权频繁项集挖掘算法,但是其只在文本分类中取得较好的效果。与普通文本相比,网页有其自身特点,如网页含有丰富的标签,结构信息更明确,使得网页分类的难度更大,需要针对其特点进一步研究。总之,将上述方法应用于网页分类时存在以下两点不足:(1)把网页当成纯文本来处理,没有考虑网页的标签信息;(2)对于关联规则中的项,要么忽略了特征词的权重,要么仅以词频作为其值,没有体现特征词在网页中所处位置的影响。

为此,本文提出基于特征词复合权重的关联网页分类方法。首先根据网页标签信息所体现的位置特征、词频等将网页表示成一个三元组集合,然后在其基础上利用基于位置特征的词频加权方法计算特征词的复合权重,并用该权重对关

到稿日期:2010-04-01 返修日期:2010-07-22 本文受国家自然科学基金(10972027)资助。

兰 均(1986-),男,硕士生,主要研究方向为数据挖掘、空间数据库,E-mail:lj508508@163.com;施化吉(1964-),男,教授,主要研究方向为数据挖掘、信息安全、空间数据库;李星毅(1969-),男,博士,副教授,主要研究方向为数据挖掘、空间数据库、交通信息系统与控制;徐 敏(1978-),女,硕士,讲师,主要研究方向为数据挖掘、神经网络。

联规则中的项进行加权,最后挖掘带复合权重的频繁项集,建立分类规则,对网页进行分类。

## 2 网页预处理

### 2.1 网页的表示

与纯文本不同,网页基于 HTML 语法规则,是一种半结构化数据。网页的结构特征由超文本标签来表达,标签域中的信息表达网页内容的能力是有差别的,因而对网页分类具有不同的指导性。如网页的标题文本含有“图书”,就表明该网页的类别与图书相关。所以分类时不仅要考虑网页中的特征词信息,而且应该注重特征词所出现的位置的影响。其中位置主要包括 Title, META, H1-H6, strong, A 和 Plain Text 等几种。

通常,Title, META 信息是对网页主题的直接描述,表明网页的中心内容,大、小标题 H1-H6 对概括和强调网页的整体和局部内容起关键作用,体现网页的内容结构;加强文本 strong 则是对比较重要文档内容的一种提示,具有很强的类别归属信息。相比之下, A 所含的锚文本提供了与网页主题相关的额外信息,具有一定的类别指示性,但较前者次之,而又比普通文本 Plain Text 要强。因此,根据不同位置的特征词所包含类别信息的差异程度<sup>[10]</sup>,本文按表 1 所述将位置分为 3 类。

表 1 位置类别及与之相对应的 HTML 标签

位置类别 $p$	HTML Tags
1	Title, strong, H1-H6, META
2	A
3	None of the above

**定义 1**(网页的三元组集合表示) 网页  $d$  可以表示成三元组集合  $R = \{(c_i, p, f_{a,p}) | c_i \in S, p \in P, f_{a,p} \in Z\}$  形式,其中  $c_i$  表示网页中的特征词,  $p$  表示  $c_i$  在网页中所处的位置,  $f_{a,p}$  表示  $c_i$  出现在位置  $p$  上的词频,  $S$  表示该网页的特征词集合,  $P$  表示位置集合,  $Z$  表示整数集。

三元组集合表示是针对网页特点的一种新方法,充分考虑到网页自身的特点,区别对待不同位置的特征词,体现了位置特征对分类的重要性。

### 2.2 词频统计

传统的文本分类方法在词频统计时,往往忽视了词语的语义关系,如多词同义等情形。一方面造成特征词的维数偏大,影响分类的效率;另一方面使分类的准确度降低。因此,本文提出基于语义的词频统计方法,即在网页的三元组表示下,求出各个特征词在语义下的词频,最终得到网页的基于语义的三元组集合表示形式。

**定义 2** 网页  $d$  中出现在位置  $p$  的特征词  $c$  在语义下的词频  $f_{c,p}$  为

$$f_{c,p} = \sum_{c_i \in U_c} f_{a,p} \quad (1)$$

式中,  $U_c = \{c, c_1, c_2, \dots, c_i\}$  表示特征词  $c$  的同义词集合。

具体步骤如下:首先根据位置类别抽取网页中的文本并分词,得到表示该网页的三元组集合;其次借助知网<sup>[11]</sup> 扩展出各个特征词的同义词,扫描三元组,将具有相同位置特征和词义的三元组合并且词频累加;最后得到网页在语义下的三元组集合。

### 2.3 特征词的复合权重

特征词的权重综合反映了该词对标识网页内容的贡献程度和网页内容之间的区分能力。基于网页自身的特点,网页中的特征词的权重取决于 3 个因素:(1)特征词在网页中的词频;(2)特征词在网页中所处的位置;(3)网页集中含有该特征词的网页数。

**定义 3**(加权词频) 定义特征词  $c$  的加权词频  $tf_c$  为

$$tf_c = \sum_{p=1}^3 (\lambda(p) \times f_{c,p}) \quad (2)$$

式中,  $f_{c,p}$  表示出现在位置  $p$  上的特征词  $c$  在语义下的词频,  $\lambda$  为位置贡献度函数。本文根据实际经验并参考前人的研究成果<sup>[10]</sup>,认为位置类别 1 的信息最能反映页面内容,应赋予最高权值,位置类别 2 的信息较能反映页面内容,赋予次高权值,具体赋值如下:

设  $p$  为位置类别特征,则

$$\lambda(p) = \begin{cases} 4, & p=1 \\ 2, & p=2 \\ 1, & p=3 \end{cases} \quad (3)$$

为此,本文提出了特征词  $c$  的复合权重  $w_c$  的计算公式

$$w_c = tf_c \times idf_c = tf_c \times \log(N/N_c + 0.01) \quad (4)$$

式中,  $tf_c$  是利用式(2)基于位置特征的词频加权方法计算的特征词  $c$  的加权词频,  $idf_c$  是反文档频率,  $N$  表示网页总数,  $N_c$  表示含有特征词  $c$  的网页数。

经过式(4)计算出各个特征词的复合权重后,为了方便后文的分类计算,将网页表示成一个向量形式,即  $V = \{(c_1, w_{c_1}), (c_2, w_{c_2}), (c_3, w_{c_3}), \dots, (c_n, w_{c_n})\}$ ,其中  $c_i$  表示特征词,  $w_{c_i}$  表示  $c_i$  的复合权重。

## 3 基于特征词复合权重的关联网页分类

在传统的关联分类方法中,用以构造关联规则的项仅是网页中的特征词,往往忽略了特征词的权重大小,或仅以词频作为其值,没有考虑到网页自身的特点。为此,本文提出的网页分类方法对其进行了考虑。首先利用网页丰富的标签信息计算特征词的复合权重,并以此对关联规则中的项进行加权,然后挖掘带复合权重的频繁项集,建立分类规则,最后对网页进行分类。

**定义 4** 已知集合  $X = \{(c_1, w_{c_1}), (c_2, w_{c_2}), \dots, (c_i, w_{c_i})\}$ ,则二元组  $(c_i, w_{c_i})$  叫做带复合权重的项(也叫加权项);  $X$  叫做加权项集。对于两个加权项  $(c_i, w_{c_i}), (c_i, w'_{c_i})$ ,如果  $w_{c_i} > w'_{c_i}$ ,则称  $(c_i, w_{c_i})$  是  $(c_i, w'_{c_i})$  的超集。

**定义 5** 已知加权项集  $X = \{(c_1, w_{c_1}), (c_2, w_{c_2}), \dots, (c_i, w_{c_i})\}$ ,最小支持度阈值为  $\epsilon$ ,则  $X$  的支持度为加权项集中最小的项的复合权重,记作  $support(X) = \text{Min}(w_{c_1}, w_{c_2}, \dots, w_{c_i})$ 。对于加权项集  $X$ ,如果  $support(X) > \epsilon$ ,则加权项集  $X$  是频繁的。

基于特征词复合权重的关联网页分类方法,分两步实现:(1)从样本网页中挖掘出所有频繁加权项集;(2)根据频繁加权项集产生所有精确分类规则,对网页进行分类。针对第一步,本文对 Apriori 算法<sup>[12]</sup>进行了改进,提出了适合本文的频繁加权项集挖掘算法 E-Apriori。

与经典 Apriori 算法类似,本文提出的 E-Apriori 算法也需要挖掘出频繁项集,但有其特点。首先 Apriori 算法主要是针对数据库,挖掘项与项之间的关系,而 E-Apriori 算法是以

网页为事务,网页中的词语为项,挖掘项与网页类别之间的关系;其次对于项的支持度,Apriori 算法是以统计项的出现频率来度量,而 E-Apriori 算法考虑到项的词频及位置特征的影响,引入了项的复合权重,并以该权重为基础计算项的支持度。总之,E-Apriori 算法是一种扩展了的 Apriori 算法,专门针对网页进行关联分类。

### 算法 E-Apriori

输入:某一类网页的训练网页集  $D$  和最小支持度阈值  $\epsilon$

输出:该类网页的频繁加权项集  $L$

- (1) For each  $d \in D$
- (2) { 对网页  $d$  进行特征抽取、词频统计等,得到其基于语义的三元组集合形式  $R$ ;
- (3) 利用基于位置特征的词频加权方法,由  $R$  得到加权项集  $T_d = \{(c_i, w_{\alpha_i}) | c_i \in S_d\}$ ;
- (4)  $add(T_d, T)$ ; //将  $T_d$  添加到集合  $T$  中
- (5) }
- (6)  $C = \{c_i | (c_i, w_{\alpha_i}) \in T\}$ ; //由加权项集得到对应的项集
- (7) For each  $c_i \in C$
- (8)  $w_{\alpha_i} = \frac{\sum_{d_i} K_{d_i}(c_i)}{|d_i | c_i \in d_i |}$ ; //计算各项的支持度,  $K_{d_i}(c_i)$  表示  $c_i$  在网页  $d_i$  中的复合权重
- (9)  $L_1 = \{(c_i, w_{\alpha_i}) | w_{\alpha_i} > \epsilon\}$ ; //  $L_1$  即是频繁 1\_项加权项集
- (10) for( $k=2; L_{k-1} \neq \text{null}; k++$ )
- (11) { 由  $L_{k-1}$  得到  $C_{k-1} = \{c_i | (c_i, w_{\alpha_i}) \in L_{k-1}\}$ ; //由频繁加权项集得到其对应的项集
- (12)  $C_k = C_{k-1} \times C_{k-1}$ ; //根据 Apriori 算法,  $k-1$  项集连接产生  $k$  项集
- (13) 对于  $\forall X. C_i \in C_k$
- (14) { For each  $c_i \in X$
- (15)  $w_{\alpha_i} = \frac{\sum_{d_i} K_{d_i}(c_i)}{|d_i | X \in d_i |}$ ; //重新再次计算每个项权重
- (16)  $L_k = \{(c_i, w_{\alpha_i}) | (c_i, w_{\alpha_i}) \in X, Support(X) > \epsilon\}$ ; //产生频繁  $k$  项加权项集
- (17) }
- (18) return  $L = \{L_i | i=1, 2, \dots, k\}$ .

按照 E-Apriori 算法产生的频繁加权项集就是初始的网页分类规则。当以一定的置信度阈值选择规则后,就组成了网页分类器。

通过构造的分类器,可以实现网页分类。但是通常分类器中的规则较多,分类时若依次匹配规则,效率较低。为此,根据前面定义的超集的性质,本文在分类时采用分类规则树来存储和查找规则。分类规则树中任意节点的子节点是其超集,匹配时若该节点处不匹配,则其子节点可以忽略匹配,进而减少了匹配次数。具体的分类步骤是:首先对待分类网页进行信息抽取、分词,计算出特征词的复合权重;其次利用分类规则树查找、匹配规则,并对该网页进行分类。

## 4 实验研究

### 4.1 实验数据及处理

所需的数据集是从互联网上收集的新闻文本,经过人工整理分成政治、教育、经济、军事、交通、娱乐、艺术和计算机 8 类,其中训练网页样本 1200 篇,每类 150 篇;测试网页 400 篇。分词技术采用中科院 ICTCLAS 中文分词系统。

数据处理:(1)利用 Tidy 工具清洗网页,纠正错误标签,然后抽取不同位置的文本信息;(2)利用 ICTCLAS 分词系统

对文本分词,并计算出在语义条件下不同位置的特征词词频,然后根据基于位置特征的词频加权方法计算特征词在网页中的复合权重;(3)用求出的复合权重对关联规则中的项进行加权,然后利用 E-Apriori 算法挖掘频繁加权项集,建立分类规则,对网页进行分类。

### 4.2 分类性能评价

在实验中分类的性能评价指标采用正确率、召回率和 F1 测试值,其计算公式如下:

$$\text{正确率 (precision)} = \frac{\text{事实属于此类且被分类正确的网页数}}{\text{被判为此类的网页数}}$$

$$\text{召回率 (recall)} = \frac{\text{事实属于此类且被分类正确的网页数}}{\text{属于此类的总网页数}}$$

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

### 4.3 实验结果

为验证本文方法的有效性,本文实现了所提出的基于特征词复合权重的关联网页分类方法(MW\_ARC),并将其与传统的不带权重的关联分类方法(N\_ARC)以及仅带词频权重的关联分类方法(CP\_ARC)进行分类效果比较。最小支持度阈值设定为 12,置信度阈值设定为 50%。3 种方法的正确率、召回率和 F1 测试值对比如图 1—图 3 所示。

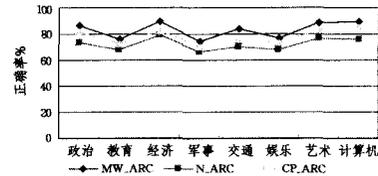


图1 正确率对比

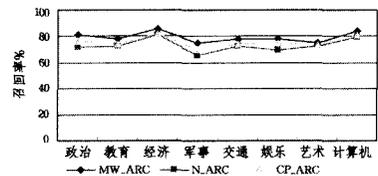


图2 召回率对比

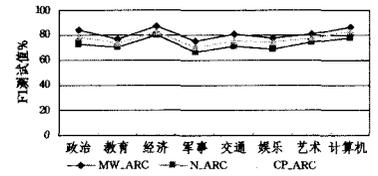


图3 F1测试值对比

从图中可以看出,带词频权重的关联分类方法(CP\_ARC)比不带权重的关联分类方法(N\_ARC)的分类效果好,说明仅仅以网页中的特征词作为关联规则中的项是不够的,需要考虑特征词权重对分类的影响,如用词频来量化权重;而基于特征词复合权重的关联网页分类(MW\_ARC)总体上又比 CP\_ARC 具有更高的分类精度,表明特征词权重的计算方式对分类准确率的提高有一定影响,说明网页不同于一般纯文本,除词频外,利用特征词在网页中的位置特征来综合计算特征词的权重能明显提高网页分类的准确度。

**结束语** 针对网页分类中现有关联分类方法的不足,本文提出了基于特征词复合权重的关联网页分类方法。实验结果表明,考虑网页的标签信息,通过基于位置特征的词频加权方法计算特征词的复合权重,并以此对关联规则中的项加权,

可以提高网页分类的准确率。而且,利用分类规则树存储分类规则,可以使分类效率得到改善。总之,本文所提出的方法一定程度上很好地弥补了现有的关联分类应用于网页分类时在分类精度上的不足,是一种比较好的网页分类方法。

### 参考文献

[1] 孙建涛,沈抖,陆玉昌,等. 网页分类技术[J]. 清华大学学报:自然科学版,2004,44(4):65-68  
 [2] 马金娜,田大纲. 基于支持向量机的中文文本自动分类研究[J]. 系统工程与电子技术,2007,29(3):475-478  
 [3] 范焱,郑诚,王清毅,等. 用 Naive Bayes 方法协调分类 Web 网页[J]. 软件学报,2001,12(9):1386-1392  
 [4] 印鉴,谭焕云. 基于  $\chi^2$  统计量的 kNN 文本分类算法[J]. 小型微型计算机系统,2007,28(6):1094-1097  
 [5] Liu Bing, Hsu W, Ma Yi-ming. Integrating Classification and Association Rule Mining[C]// ACM International Conference on Knowledge Discovery and Data Mining(SIGKDD'98). 1998: 80-86  
 [6] Li Wen-min, Han Jia-wei, Pei Jian. CMAR: Accurate and Effi-

cient Classification Based on Multiple Class association Rules [C]//First IEEE International Conference on Data Mining(ICDM'01). 2001:396-376  
 [7] Guo Yu-qin, Yuan Fang, Liu Hai-bo. Text categorization based on fuzzy classification rules tree[J]. Journal of Southeast University(English Edition), 2008,24(3):339-342  
 [8] 王元珍,钱铁云,冯小年. 基于关联规则挖掘的中文文本自动分类[J]. 小型微型计算机系统,2005,26(8):1380-1383  
 [9] 邱江涛,唐常杰,乔少杰,等. 基于加权频繁项集的文本分类规则挖掘[J]. 四川大学学报:工程科学版,2008,40(6):110-114  
 [10] Cutler M, Shi Yung-ming, Meng Wei-yi. Using the Structure of HTML Documents to Improve Retrieval[C]//Proceeding of the USENIX Symposium on Internet Technologies and Systems Monterey, California, 1997:22-33  
 [11] 董振东,董强. 知网[EB/OL]. <http://www.keenage.com>, 2009  
 [12] Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Databases[C]// Proceeding of the 1993 ACM SIGMOD Conference, Washington DC, USA, 1993:207-216

(上接第 152 页)

群签名的验证方程为:

$$g^s r^r \text{ mod } q \equiv (Y_G)^{h(m)} \pmod{p} \quad (5)$$

**定理 2** 单签名的正确性可以通过式(4)验证。

证明:由单签名的产生过程知,

$$g^{s_i} r_i^r \equiv g^{c_i \gamma_i h(m) - k_i r} \cdot g^{k_i r} \equiv ((\delta_i)^{c_i})^{h(m)} \equiv (s_i')^{h(m)} \pmod{p}$$

因此,单签名满足式(4)。证毕。

**定理 3** 群签名的正确性可以通过式(5)验证。

证明:由签名过程及定理 2 知,

$$g^s r^r \equiv g^{\sum s_i} (\prod r_i)^r \equiv g^{h(m) \sum c_i \gamma_i - r \sum k_i} \cdot g^{\sum k_i r} \\ \equiv (g^{\sum c_i \gamma_i})^{h(m)} \equiv (Y_G)^{h(m)} \pmod{p}$$

因此,群签名满足式(5)。证毕。

## 5 分析与讨论

群签名方案具有以下特性:

(1) “特权集”与“门限”特性

由秘密密钥分割过程及定理 1 的证明易见,如果不足  $t$  个用户参加签名,或者即使有  $t$  个以上用户参加签名但不满足特权条件,则不能恢复群秘密密钥,从而无法生成有效的群签名。

(2) 秘密共享的“理想”特性

对群秘密密钥分割时,群中任一成员只需保存一个秘密密钥碎片信息,秘密共享的信息率为 1。

(3) 签名的不可伪造性

在基本 ElGamal 签名的安全性假设下,首先除 KAC 之外,任何成员不可能伪造其他成员的单签名,从而任意  $t-1$  个成员不可能通过伪造其他成员的单签名来伪造群签名;其次,由定理 1 知,任意  $t-1$  个成员不能得到有关群秘密密钥的任何信息,从而无法直接伪造群签名。因此,本方案可以抵抗任意  $t-1$  个成员的合谋攻击。

(4) 验证的匿名性与身份的可追查性

验证者 V 利用群公钥可以验证签名的有效性,单签名参与者的身份信息对 V 来说是无法确定的,因此具有签名验证的匿名性;SC 知道签名者身份,如果得到许可,由 SC 追查签名者身份是平凡的。

说明:

(1) 本方案中,将“特权集”的思想嵌入群秘密密钥碎片的产生过程。在实际应用中,为保证方案的安全性,KAC 在进行秘密分割时需要进行检验,以保证秘密共享的完善性。

(2) 与文献[6]中的特权集方案相比较,本方案的优点是具有秘密共享的理想性,缺点是通信代价有所增加(这也是需要进一步解决的问题)。除此之外,二者签名长度、签名和验证的代价都相当。

**结束语** 本文将“分部秘密共享”技术和“特权集”思想结合在一起,提出了一种存在特权集的门限群签名方案。基于代数中的基本结论,证明了秘密共享方案的安全性。本文方案除了满足一般群签名体制的基本性质之外,还具有“特权集”与“门限”特性、秘密共享的理想性、签名的不可伪造性、验证的匿名性与身份的可追查性等良好特性,在成员存储能力和计算能力受限的场合具有重要的应用价值。

### 参考文献

[1] Desmedt Y, Frankel Y. Shared generation of authenticators and signatures[C]// Proceedings of Cryptology-Crypto'91. Berlin: Springer-Verlag, 1991:457-469  
 [2] Harn L. Group-oriented  $(t, n)$ -threshold digital signature scheme based on discrete logarithms[J]. IEEE Proceedings of Computers and Digital Techniques, 1994, 141(5):307-313  
 [3] 徐秋亮. 改进门限 RSA 数字签名体制[J]. 计算机学报, 2000, 23(5):449-453  
 [4] 蒋翰,徐秋亮,周永彬. 基于 RSA 密码体制的门限代理签名[J]. 计算机学报, 2007, 30(2):241-247  
 [5] 马春波,何大可. 矢量空间秘密共享群签名方案[J]. 电子学报, 2005, 33(2):294-296  
 [6] 陈伟东,冯登国. 一类存在特权集的门限群签名方案[J]. 软件学报, 2005, 16(7):1289-1295  
 [7] 王天芹. 存在特权集的门限代理群签名方案[J]. 计算机应用研究, 2008, 25(7):2146-2147  
 [8] Farras O, Marti-Farre J, Padro C. Ideal multipartite secret sharing schemes[C]// Lecture Notes in Computer Science 4515. Berlin: Springer-Verlag, 2007:448-465