

# Internet 上的文本数据挖掘

Text Mining on the Internet

王伟强 高文 段立娟

(中国科学院计算技术研究所 北京 100080)

**Abstract** The booming growth of the Internet has made text mining on it a promising research field in practice. The paper summarily introduces some aspects about it, which involve some potential applications, some techniques used and some present systems.

**Keywords** Text mining, Word sequence, Feature subset

## 1 引言

今天 Internet 已不再是科学家及工程师们独享的通讯工具,已成为数字化时代的世界性图书馆,变成了为各行各业的人们交流思想、获取信息的便利手段。人们在使用 WWW 浏览服务时,检索、获取最多的信息数据就是文本数据。但这种在巨大的 Internet 信息仓库中占信息比重最大的数据类型却缺乏结构化、组织的规整性,并且随意地散布在这个网络的各个角落,还降低了人们对丰富的信息资源的利用效率。数据挖掘是人们对一些巨型数据库中的数据进行分析、使用感到力不从心时而发展出的一门新的技术,它将人工智能技术与数据库技术紧密结合起来,让计算机帮助人们从庞大的数据中智能地、自动地抽取有价值的知识模式,以满足人们不同应用的需要。当数据挖掘的对象完全由文本这种数据类型组成时,这个过程就称文本数据挖掘。Internet 上缺乏结构化、组织规整性的超巨量文本数据自然让人们想到运用文本数据挖掘技术来帮助提高人们在 Internet 上检索信息、利用信息的效率。

事实上,文本数据挖掘在 Internet 上存在许多潜在应用,这里举出其中的几个例子。

(a)在搜索引擎上对文档进行自动分类。搜索引擎可以帮助人们对 Internet 上庞大而杂乱无章的数据进行电子索引,大大地提高人们对 Internet 上数据的使用效率。搜索引擎程序周期性地向它能识别的每一个站点发送一段有人称为网上爬虫(Web crawler)或也有人称为索引机器人的程序,可将各站点上的网页下载下来,搜索引擎可自动地从这些网页上抽取描述它们的索引信息,将该信息连同相应网页的地址 URL 一并存入搜索引擎的数据库中。在此基础上人们可利

用数据挖掘技术对该数据库中的信息进行进一步的自动化整理,自动生成便于用户使用的网页分类系统,即现在我们见到的用超连接组织 Internet 上文档的方式,从而大大降低了在搜索引擎上为组织、整理 Internet 上文档所需耗费的人力资源。

(b)帮助寻找用户感兴趣的新闻或其它信息。许多 Internet 用户在日常生活中都有阅读网上新闻的习惯,这样就可以设计一个电子新闻过滤系统,利用文本机器学习建立起该用户的趣向模型,每当用户进入一份电子报纸的网页时,该系统就会根据学习所得的模型对其中的每一篇文章按与该用户兴趣的接近程度进行打分排序,使用户最先看到的是他最感兴趣的新闻。类似的思想还可以帮助用户选择他所感兴趣的超连接。

(c)对用户的检索结果实现更友好的人机接口。当人们利用现有的搜索引擎来检索用户希望获得的信息时,搜索引擎会根据用户输入的关键字提供一个查询结果的线性表。通常的情况下,这个表是很大的且其中不可避免地存在很多的无关信息,这时若能对查询的结果进行一下分析聚类,然后以一种超连接组织的层次方式提交给用户,则会给用户筛选查询结果的信息带来极大的方便。

## 2 Internet 上文本数据挖掘技术介绍

### 2.1 文本数据挖掘的一般处理过程

Internet 上文本数据挖掘的一般处理过程可用图 1 来概括描述。首先对挖掘对象建立其特征表示,在 Internet 上的文本数据挖掘对象通常是一组 html 格式的文档集,这样的挖掘对象缺乏象关系数据库中数据的组织规整性,因此要将这些文档转化成一种类似关系数据库中记录的较规整且能反映文档内容特征的

表示,一般采用一个文档特征向量,但在目前所采用的文档表示方法中,存在的一个共同的不合人意的地方是文档特征向量具有惊人的维数,使得特征子集的选取成为 Internet 上文本数据挖掘过程中的必不可少的一个环节。在完成文档特征向量维数的缩减后,便可利用机器学习的各种方法来提取面向特定应用目的的知识模式。最后对获取的知识模型进行质量评价,若评价的结果满足一定的要求,则存储该知识模式,否则返回到以前的某个环节分析改进后进行新一轮的挖掘工作。下面对各个环节常用的一些方法进行分别的介绍。

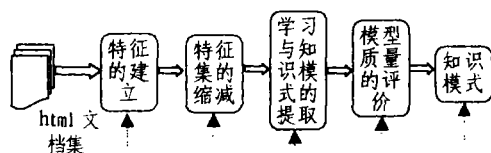


图1 Internet上文本数据挖掘的一般处理过程

## 2.2 文档的表示

在对文本数据进行学习前,需要把它表示成一种特定的形式,在信息检索和文本机器学习中频繁被采用的文档表示方法叫做 TFIDF 向量表示法,TFIDF 向量反映了训练文档集的单字空间,它的每个向量分量对应一个单字,分量的大小  $d(i)$  为  $TF(W_i, Doc)$  与  $IDF(W_i) = \log D/DF(W_i)$  的乘积,其中  $TF(W_i, Doc)$  为单字  $W_i$  在文档  $Doc$  中的出现频度, $D$  为总文档数, $DF(W_i)$  为单字  $W_i$  在其中出现至少一次的文档的数目。我们可以这样理解  $d(i)$ ,  $d(i)$  刻画了单字  $W_i$  区分文档内容属性的能力,当一个单字在文档集中出现的范围越广,说明它区分文档属性的能力越低。另一方面,它在一特定的文档中出现的频度越高,说明它在区分该文档内容属性方面的能力越强。它属于一种文档的词集(bag-of-words)表示法,即所有的词从文档中抽取出来,而放弃考虑词间的次序以及文本的结构。此外许多采用词集文档表示的系统,或者用一个布尔值作为文档向量的分量来表示特定的单字在文档中是否出现,或者用词在特定文档中的出现频率作为文档向量的分量。有的系统还使用了象词的位置等这样的其它信息。

一个系统中具体采用什么样的文档表示,与该系统进行数据挖掘的目的有关。例如在目的是根据一个用户访问过的 WEB 文档集来学习、预测用户将要检索的超连接的个人导游器系统<sup>[2,3]</sup>中,系统在抽取文档集的特征时选取了词串(word sequence, 又称 n-grams)作为特征来表示文档。下面给出了以词串作为特征的文档表示的文档集特征提取算法,该算法主要

是针对英语语言环境设计的<sup>[1,10]</sup>。算法中所涉及的排除词表(stop-list)与特定的语言有关,它一般由特定语言中出现频率最高但含义虚泛的词来构成,例如英语中的 a, the, for, each at, will, be 等,汉语中的“的、地、这、那、虽然、但是”等。首先算法消除了那些出现在排除词表中的单字选做特征的可能性。然后又排除那些在文档集中出现频度很低的单字作为特征,这一点可通过对文档集进行单字频度统计并选择一个合适的门限值来做到,例如选取  $MinNGramOcc = 5$ 。这个特征抽取过程要对文档集进行多遍的扫描,扫描的遍数由人设定的可作为特征的最大词串长度来决定,例如选取  $MaxNGramSize = 5$ 。第一遍扫描使所有不存在于排除词表中且拥有足够出现频度的单字被收入到特征词串表集  $LargeNGramSet$  中。对于长度为 2 到  $MaxNGramSize$  的词串利用多个筛选标准来抽取。在每一遍的扫描中,所有的文档均利用一个窗口队列  $NGramQueue$  来逐个单字地进行检查以获得一个个的词串,每个文档中的符号若要进入该窗口必须满足它是一个正确的单字(而非一个数字或特殊的符号),它不存在于排除词表中且属于当前的字串集  $LargeNGramSet$  中;否则该窗口将被清空复位。

## 文档集特征提取算法

变量说明:

输入:  $MinNGramOcc$ : 被选作文档特征的词串的最小的出现次数  
 $MaxNGramSize$ : 词串的最大长度  
 $StopWordSet$ : 排除词表(跟特定的语言有关)  
 $DocVec$ : 文档向量  
 $SymVec$ : 文档中的符号向量  
中间:  $Sym$ : 文档中的符号(可能值的类别为: 单字、数字、标点符号)  
 $CandNGramMap$ : 映射候选词串为当前它的出现次数  
 $NGramQueue$ : 存储当前长度为  $NGramSize$  的词串  
输出:  $LargeNGramSet$ : 满足条件可作为文档特征的最大字串集

算法描述:

```
(1)  $LargeNGramSet = \text{all single words in doc Vec, not in Stop WordSet and occurring} \geq MinNGramOcc$ ;
(2) for  $NGramSize = 2$  to  $MaxNGramSize$  do
(3) {
(4)  $CandNGramMap = []$ ;
(5) For  $SymVec = DocVec[1]$  to  $DocVec[|DocVec|]$  do
(6) {
(7)  $NGramQueue = []$ ;
(8) For  $Sym = SymVec[1]$  to  $SymVec[|SymVec|]$  do
(9) {
(10) if ( $TypeOf(Sym) == \text{word}$ ) {
(11) if ( $Sym \text{ not in } StopWordSet$ ) {
(12) if ( $(|NGramQueue| + 1) == NGramSize$ ) {
(13) if ( $Concatenated(NGramQueue)$  in  $LargeNGramSet$ ) {
(14) // 该词已在  $LargeNGramSet$  中出现
(15)  $NGramQueue.Push(Sym)$ ;
(16)  $CandNGramMap[Concatenated(NGramQueue)]++$ ;
(17)  $NGramQueue.Pop()$ ;
(18) }
(19) else {
(20)  $NGramQueue.Push(Sym)$ ;
(21)  $NGramQueue.Pop()$ ;
```

```

(22) }
(23) }
(24) else // |NGramQueue|+1<NGramSize
(25)     NGramQueue.Push(Sym);
(26) } // Sym in LargeNGramSet
(27) else NGramQueue=[];
(28) } // Sym not in StopWordSet
(29) else NGramQueue=[];
(30) } // TypeOf(Sym)==word
(31) else NGramQueue=[];
(32) } // for sym
(33) } // for symvec
(34) LargeNGramSet += { NGram: CandNGramMap
    [NGram]>+MinNGramOcc};
(35);
(36) return LargeNGramSet;

```

### 2.3 特征子集的选取

使用前面提到的词集表示法来表示 WWW 上的待学习的文档时,表示文档的特征向量会达到数十万维的大小。有人曾利用上面提到的文档集特征提取算法对 Yahoo 上 49600 个文档提取作为特征的词串(最大长度为 5),最后得到 320000 个特征词串<sup>[1]</sup>。如此高维的特征对将进行的分类机器学习未必全是重要、有益的,而且高维的特征可能会大大增加机器的学习时间而仅产生与小得多的特征子集相关的学习分类结果,因此在对 WWW 上的文档进行机器学习时特征子集的选取便显得非常重要。

虽然在机器学习中有许多特征子集选取算法,但在面对这种超高维的特征集时变得不再适用。目前对 WWW 文档特征所采用的特征子集选取算法一般是构造一个评价函数,对特征集中的每个特征进行独立的评估,这样每个特征都获得一个评估分,然后对所有的特征按照其评估分大小进行排序,选取预定数目的最佳特征作为结果的特征子集。所以,选取多少个最佳特征以及采用什么评价函数都需要针对一个具体的问题通过试验来决定。

一些已被采用的评估函数有信息增益(Information Gain)、期望交叉熵<sup>[11]</sup>(Expected Cross Entropy)、互信息<sup>[5]</sup>(Mutual Information)、文本证据权<sup>[1]</sup>(the Weight of Evidence for Text)、几率比<sup>[12]</sup>(odds ratio)、词频<sup>[5]</sup>(Word Frequency)等。下面列出了它们在 Internet 上进行文本数据挖掘时所采用的一些数学表示。

$$InfGain(F) = P(W) \sum_i P(C_i|W) \log \frac{P(C_i|W)}{P(C_i)} + P(W) \sum_i P(C_i|\bar{W}) \log \frac{P(C_i|\bar{W})}{P(C_i)} \quad (1)$$

$$CrossEntryTxt(F) = P(W) \sum_i P(C_i|W) \log \frac{P(C_i|W)}{P(C_i)} \quad (2)$$

$$MutualInfoTtxt(F) = \sum_i P(C_i) \log \frac{P(W|C_i)}{P(W)} \quad (3)$$

$$WeightofEvidTtxt(F) = P(W) \sum_i P(C_i)$$

$$\left| \log \frac{P(C_i|W)(1-P(C_i))}{P(C_i)(1-P(C_i|W))} \right| \quad (4)$$

$$OddsRatio(F) = \log \frac{P(W|pos)(1-P(W|neg))}{P(W|neg)(1-P(W|pos))} \quad (5)$$

$$Freq(F) = TF(W) \quad (6)$$

其中  $F$  为对应于单字  $W$  的特征,  $P(W)$  为单字  $W$  出现的概率,  $\bar{W}$  意味着单字  $W$  并不出现,  $P(C_i)$  为第  $i$  类值的出现概率,  $P(C_i|W)$  为当单字  $W$  出现时属于第  $i$  类的条件概率,  $P(W|pos)$  为在类  $pos$  中单字  $W$  出现的条件概率,  $P(W|neg)$  为在类  $neg$  中单字  $W$  出现的条件概率,  $TF(W)$  为单字在文档集中出现的次数。

### 2.4 文档进行分类的常用算法

对 Internet 上的文本数据进行学习的一个重要应用就是根据一个文档中所包含的文本数据的特征给出该文档的类别,使它进入到合适的主题范围中去,以方便对它的查阅检索。这里介绍一下朴素贝叶斯分类算法<sup>[13,14]</sup>(Naive Bayesian classifier)和  $k$ -最近邻参照分类算法<sup>[1]</sup>( $k$ -Nearest Neighbor)在文档分类上的具体运用。

A. 利用朴素贝叶斯分类算法进行文档分类 若文档采用  $Df$  向量表示法,即文档向量的分量为一个布尔值,0 表示相应的单字在该文档中未出现,1 表示出现,则采用该表示方法的文档  $Doc$  属于  $C$  类文档的概率为:

$$P(C|Doc) = \frac{P(C) \prod_{F_i \in V} P(Doc(F_i)|C)}{\sum_i P(C_i) \prod_{F_i \in V} P(Doc(F_i)|C_i)} \quad (7)$$

$$P(Doc(F_i)|C) = \frac{1 + N(Doc(F_i), C)}{2 + |D_c|}$$

其中  $P(Doc(F_i)|C)$  是对在  $C$  类文档中特征  $F_i$  出现的条件概率的拉普拉斯概率估计,  $N(Doc(F_i), C)$  是  $C$  类文档中特征  $F_i$  出现的文档数,  $|D_c|$  为  $C$  类文档所包含的文档的数目。

若文档采用  $Tf$  向量表示法,即文档向量的分量为相应的单字在该文档中出现的频度,则采用该表示方法的文档  $Doc$  属于  $C$  类文档的概率为:

$$P(C|Doc) = \frac{P(C) \prod_{F_i \in V} P(F_i|C)^{TF(F_i, Doc)}}{\sum_i P(C_i) \prod_{F_i \in V} P(F_i|C_i)^{TF(F_i, Doc)}} \quad (8)$$

$$P(F_i|C) = \frac{1 + TF(F_i, C)}{|V| + \sum_i TF(F_i, C)}$$

其中  $P(C)$  为一个文档属于  $C$  类的概率,  $P(F_i|C)$  是对在  $C$  类文档中特征  $F_i$  出现的条件概率的拉普拉斯概率估计,  $TF(F_i, C)$  是  $C$  类文档中特征  $F_i$  出现的频度,  $|V|$  为单字辞典集的大小,等于文档表示中所包含的不同特征的总数目,  $TF(F_i, Doc)$  是在文档  $Doc$  中特征  $F_i$  出现的频度。由于在单字辞典集  $|V|$  中许多的特征  $F_i$  均不出现于文档  $Doc$  中,从而  $TF(F_i, Doc) = 0$ , 所

以可将式(8)中第一个公式变作:

$$P(C|Doc) = \frac{P(C) \prod_{F_j \in Doc} P(F_j|C)^{TF(F_j, Doc)}}{\sum_i P(C_i) \prod_{F_j \in Doc} P(F_j|C_i)^{TF(F_j, Doc)}} \quad (9)$$

B. 利用 k-最近邻参照分类算法进行文档分类  
k-最近邻参照分类算法将对一个文档的所属类别范畴的预测建立在与之最为相似的 k 个文档所属类别的概率分布上。文档 Doc 属于 C 类文档的概率为:

$$P(C|Doc) = \frac{\sum_{i=1}^k \text{similarity}(Doc, D_i) P(C|D_i)}{\sum_i \sum_{i=1}^k \text{similarity}(Doc, D_i) P(C_i|D_i)} \quad (10)$$

其中  $D_i$  为与文档 Doc 最近邻的 k 个文档之一,它既可以按不同的概率属于不同的类别,也可以属于唯一的一个类别  $C_k$  (这时  $P(C_i|D_i)=1$ , 当  $i=k$  时;否则,  $P(C_i|D_i)=0$ )。在信息检索中两个文档的相似程度常用表示两个向量的夹角的余弦来度量,即:

$$\cos(\bar{X}, \bar{Y}) = \frac{\sum_i X_i Y_i}{\sqrt{\sum_i X_i^2 \sum_i Y_i^2}} \quad (11)$$

## 2.5 模型质量评价

对 Internet 上的文本数据进行数据挖掘可以看作是一种机器学习的过程。在机器学习中学到的结果是某种知识模型 M, 机器学习的一个重要组成部分便是对产生的模型 M 进行评估。常用的方法有预留法(hold-out)和交叉验证法(cross-validation)。两种方法均将数据集分成训练集与测试集两部分。学习-测试循环反复执行,最后用一个平均质量来衡量模型质量的好坏。在预留法中从数据集中随机抽取预定大小的一个子集作为测试集,其余的数据作为训练集;在交叉验证法中,整个数据集按照所要进行的学习-测试循环次数分成相当数目的子集,在每次循环中,其中的一个子集作为测试集,而其它子集的并集作为训练集。

在机器学习中常用的模型质量估计方法有分类正确率(classification accuracy),查准率(precision)与查全率<sup>[15]</sup>(recall),查准率与查全率的几何平均数<sup>[16]</sup>,信息估值<sup>[17]</sup>(information score)等。它们的定义描述如下:

### A. 分类正确率

$$\begin{aligned} \text{Accuracy}(M) &= \sum_{ex} P(ex) \text{Accuracy}(M, ex) = \\ &P(\hat{C}(ex) = C(ex)) \\ \text{Accuracy}(M, ex) &= \begin{cases} 1; \hat{C}(ex) = C(ex) \\ 0; \text{其它} \end{cases} \end{aligned} \quad (12)$$

其中  $C(ex)$  为样例  $ex$  的实际类值,  $\hat{C}(ex)$  为通过模型 M 对样例  $ex$  的预测类值,  $P(ex)$  为样例  $ex$  的概率(通常为  $1/n$ ,  $n$  为样本集的大小)。

### B. 查准率与查全率

查准率定义为检索到的目

标类的样例集中所包含的属于检索正确的样例所占的比例的大小。对目标类 targetC 模型 M 的查准率可用公式(13)来估计。

$$\text{Precision}(M, t \arg etC) = P(t \arg etC | t \arg etC) \quad (13)$$

查全率定义为在一个检索结果中所包含的检索正确的对象数目占实际存在的满足查询要求的对象数目的比例的大小。对目标类 targetC 模型 M 的查全率可用公式(14)来估计。

$$\text{Recall}(M, t \arg etC) = P(t \arg etC | t \arg etC) \quad (14)$$

在式(13)(14)中  $t \arg etC$  代表实际值为目标类值,  $t \arg etC$  代表预测值为目标类值。

有时可以将两者结合起来估计模型的质量,如在信息检索中常用的 F 方法,两者在估计模型质量的相对重要性时用一个参数  $\beta$  来刻画,见公式(15)。

$$F_\beta(M, t \arg etC) = \frac{(1+\beta^2) \text{Precision}(M, t \arg etC) \text{Recall}(M, t \arg etC)}{\beta^2 \text{Precision}(M, t \arg etC) + \text{Recall}(M, t \arg etC)} \quad (15)$$

其中  $\beta$  的取值范围为  $[0, \infty)$ , 当  $\beta=0$  时,  $F_\beta$  即为查准率, 当  $\beta=\infty$  时,  $F_\beta$  即为查全率; 当  $\beta=1$  时, 查准率与查全率在估价模型质量时具有同样的重要性; 若  $\beta<1$ , 则强调查准率的作用, 若  $\beta>1$ , 则强调查全率的作用。

C. 查准率与查全率的几何平均数 文[16]中采用了查准率与查全率的几何平均数来衡量模型的质量,如公式(16)(17)所示。

$$\text{Geo Precision}(M) = \sqrt{\prod_{i=1}^{Cl_s} \text{Precision}(M, C_i)} \quad (16)$$

$$\text{Geo Recall}(M) = \sqrt{\prod_{i=1}^{Cl_s} \text{Recall}(M, C_i)} \quad (17)$$

其中  $Cl_s$  为总的类别数。

D. 信息估值 信息估值  $\text{InfoScore}(M)$  可衡量有关一个模型给出的实际类值的平均信息量,其定义如公式(18)所示。

$$\begin{aligned} \text{InfoScore}(M) &= \sum_{ex \in \text{TrainSet}} \text{InfoScore}(M, ex) \\ \text{InfoScore}(M, ex) &= \begin{cases} \log \frac{\hat{P}(C(ex))}{Pa(C(ex))}; \hat{P}(C(ex)) > Pa(C(ex)) \\ -\log \frac{1-\hat{P}(C(ex))}{1-Pa(C(ex))}; \hat{P}(C(ex)) \leq Pa(C(ex)) \end{cases} \end{aligned} \quad (18)$$

其中  $C(ex)$  是样例  $ex$  的实际类值,  $\hat{P}(C(ex))$  为对实际类值的预测概率,  $Pa(C(ex))$  为实际类值的先验概率。

## 3 Internet 上文本数据挖掘的一些系统

目前世界上的一些大学、机构和公司都致力于 In-

ternet 上文本数据挖掘系统的开发,这里对它们中的一些工作做一下介绍。

(1)WebWatcher 与 Personal WebWatcher. Web-Watcher 是由 CMU 开发的一个可安装在一个 WWW 站点上的导游器(tour guide),WebWatcher 对来访的用户的访问行为进行在线的学习,通过对站点上主页的超文本结构和以前用户浏览路径的学习,建立起一个经验模型。当一个用户进入该站点时,系统提供一个接口来启动 WebWatcher 的导游功能,它将陪伴用户进入每一个网页,同时通过对用户兴趣的分析向用户建议下一步他要访问的连接。Personal WebWatcher 是一个个人化导游器,它与 WebWatcher 的功用很相似,也由 CMU 开发。Personal WebWatcher 与 Web-Watcher 的区别在于后者是面向特定的个人而前者是面向特定的 WWW 站点。

(2)AltaVista Discovery.是由 DEC 公司开发的一个新型的桌面信息检索工具,它提供了对桌面、Internet、Usenet 数据的无缝集成。它可以基于内容在不同搜索空间进行检索,如本地盘、网络盘、Internet。可以自动对所搜索到的文档进行总结、寻找与当前网页相关联的网页,如内容相似、曾对该网页进行过引用的网页等。

**结束语** 在 Internet 迅猛发展的今天,在 Internet 上开展各种应用目的的数据挖掘研究有着良好的商业前景和实用价值。由于在 Internet 上被挖掘对象的独有特点,使它具有一套不同于其它挖掘系统的处理方法。在一般的挖掘系统中,待挖掘对象一般为一个关系数据库中的表或视图,而其中的属性均由人来直接定义以反映训练例的特征,目前数据挖掘领域的许多成果均可作为合适的挖掘手段进行尝试。在 Internet 上进行数据挖掘时,作为训练例的是一组 HTML 文档,这里文档特征的提取已成为数据挖掘过程的必要环节,提取后的挖掘对象往往是一个超高维的例子集,这种数据特点向人们提出了新的挑战,等待着人们去探索新型高效的处理方法。

最后本文提供一些含有相关信息的 WWW 站点供参考:

<http://www.kdnuggets.com/>  
<http://www.sinokdd.163.net/>  
<http://www.almaden.ibm.com/cs/people/ragraval/pubs.html#txt>  
<http://www.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www/>

## 参考文献

- 1 Mladenic D. Machine Learning on non-homogeneous, distributed text data. doctoral dissertation, University of

Ljubljana, 1998

- 2 Joachims T, et al. WebWatcher: A Tour Guide for the World Wide Web. In: Proc. of IJCAI97, August 1997
- 3 Armstrong R, et al. WebWatcher: A Learning Apprentice for the World Wide Web. In: AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, March 1995
- 4 Mladenic D. Feature subset selection in text-learning. In: Proc. of the 10th European Conf. on Machine Learning ECML98, 1998
- 5 Yang Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization. In: Proc. of the 14th Intl. Conf. on Machine Learning ICML97, 1997. 412~420
- 6 Akkas A, Giivenir H A. K Nearest Neighbor Classification on Feature Projections. In: Proc. of the 13th Intl Conf. on Machine Learning ICML96, 1996
- 7 Grobelnik M, Mladenic D. Efficient Text Categorization. In: Proc. of ECML98 Text Mining Workshop
- 8 Kamba T, et al. ANATAGONOMY: a Personalized Newspaper on the World Wide Web. Intl. Journal Human-Computer Studies 46
- 9 Mladenic D. Turning Yahoo into an Automatic Web-Page Classifier. In: Proc. of the 13th European Conf. on Artificial Intelligence ECAI'98, 1998
- 10 Mladenic D, Grobelnik M. Word Sequences as Features in Text-learning. In: Proc. of the 7th Electrotechnical and Computer Sc. Conf. ERK'98, 1998
- 11 Koller D, Sahami M. Hierarchically classifying documents using very few words. ICML97, 1997. 170~178
- 12 van Rijsbergen C J, et al. the selection of good search terms. Information Processing & Management, 1981, 17: 77~91
- 13 Joachims T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. ICML97, 1997. 143~151
- 14 Mladenic D. Personal Web Watcher: Implementation and Design. [Technical Report, IJS-DP-7472]. Oct., 1996, Available at: <http://www-ai.ijs.si/DunjaMladenic/papers/pww/>
- 15 Lewis D D. Evaluating and Optimizing Autonomous Text Classification Systems. In: Proc. of the 7th Annual Interl. ACM-SIGIR Conf. on Research and Development in Information Retrieval, Dublin
- 16 Kubat M, et al. Learning when Negative Examples abound. In: Proc. of the 9th European Conf. on Machine Learning ECML97. 1997. 146~153
- 17 Kononenko I, Bratko I. Information-Based Evaluation Criterion for Classifier's Performance. Machine Learning, Kluwer Academic Publishers. 67~80
- 18 Mladenic D, Grobelnik M. Feature selection for classification based on text hierarchy. Working Notes of Learning from Text and the Web, Conf. on Automated Learning and Discovery CONALD-98, 1998