**一、课题背景概述**

文本挖掘是一门交叉性学科,涉及数据挖掘、[**机器学习**](http://lib.csdn.net/base/machinelearning)、模式识别、人工[**智能**](http://lib.csdn.net/base/aiplanning)、统计学、计算机语言学、[**计算机网络**](http://lib.csdn.net/base/computernetworks)技术、信息学等多个领域。文本挖掘就是从大量的文档中发现隐含知识和模式的一种方法和工具,它从数据挖掘发展而来,但与传统的数据挖掘又有许多不同。文本挖掘的对象是海量、异构、分布的文档(web);文档内容是人类所使用的[**自然语言**](http://lib.csdn.net/base/nlp),缺乏计算机可理解的语义。传统数据挖掘所处理的数据是结构化的,而文档(web)都是半结构或无结构的。所以,文本挖掘面临的首要问题是如何在计算机中合理地表示文本,使之既要包含足够的信息以反映文本的特征,又不至于过于复杂使学习[**算法**](http://lib.csdn.net/base/datastructure)无法处理。在浩如烟海的网络信息中,80%的信息是以文本的形式存放的，WEB文本挖掘是WEB内容挖掘的一种重要形式。

文本的表示及其特征项的选取是文本挖掘、信息检索的一个基本问题，它把从文本中抽取出的特征词进行量化来表示文本信息。将它们从一个无结构的原始文本转化为结构化的计算机可以识别处理的信息，即对文本进行科学的抽象，建立它的数学模型，用以描述和代替文本。使计算机能够通过对这种模型的计算和操作来实现对文本的识别。由于文本是非结构化的数据,要想从大量的文本中挖掘有用的信息就必须首先将文本转化为可处理的结构化形式。目前人们通常采用向量空间模型来描述文本向量,但是如果直接用分词算法和词频统计方法得到的特征项来表示文本向量中的各个维,那么这个向量的维度将是非常的大。这种未经处理的文本矢量不仅给后续工作带来巨大的计算开销,使整个处理过程的效率非常低下,而且会损害分类、聚类算法的精确性,从而使所得到的结果很难令人满意。因此,必须对文本向量做进一步净化处理,在保证原文含义的基础上,找出对文本特征类别最具代表性的文本特征。为了解决这个问题,最有效的办法就是通过特征选择来降维。

目前有关文本表示的研究主要集中于文本表示模型的选择和特征词选择算法的选取上。用于表示文本的基本单位通常称为文本的特征或特征项。特征项必须具备一定的特性:1)特征项要能够确实标识文本内容;2)特征项具有将目标文本与其他文本相区分的能力;3)特征项的个数不能太多;4)特征项分离要比较容易实现。在中文文本中可以采用字、词或短语作为表示文本的特征项。相比较而言，词比字具有更强的表达能力，而词和短语相比，词的切分难度比短语的切分难度小得多。因此，目前大多数中文文本分类系统都采用词作为特征项，称作特征词。这些特征词作为文档的中间表示形式，用来实现文档与文档、文档与用户目标之间的相似度计算 。如果把所有的词都作为特征项，那么特征向量的维数将过于巨大，从而导致计算量太大，在这样的情况下，要完成文本分类几乎是不可能的。特征抽取的主要功能是在不损伤文本核心信息的情况下尽量减少要处理的单词数，以此来降低向量空间维数，从而简化计算，提高文本处理的速度和效率。文本特征选择对文本内容的过滤和分类、聚类处理、自动摘要以及用户兴趣模式发现、知识发现等有关方面的研究都有非常重要的影响。通常根据某个特征评估函数计算各个特征的评分值，然后按评分值对这些特征进行排序，选取若干个评分值最高的作为特征词，这就是特征抽取(Feature Selection)。

特征选取的方式有4种：(I)用映射或变换的方法把原始特征变换为较少的新特征；(2)从原始特征中挑选出一些最具代表性的特征；(3)根据专家的知识挑选最有影响的特征；(4)用数学的方法进行选取，找出最具分类信息的特征，这种方法是一种比较精确的方法，人为因素的干扰较少，尤其适合于文本自动分类挖掘系统的应用。

随着网络知识组织、[**人工智能**](http://lib.csdn.net/base/aiframework)等学科的发展,文本特征提取将向着数字化、智能化、语义化的方向深入发展,在社会知识管理方面发挥更大的作用。

**二、文本特征向量**

经典的向量空间模型(VSM: Vector Space Model)由Salton等人于60年代提出，并成功地应用于著名的SMART文本检索系统。VSM概念简单,把对文本内容的处理简化为向量空间中的向量运算,并且它以空间上的相似度表达语义的相似度,直观易懂。当文档被表示为文档空间的向量，就可以通过计算向量之间的相似性来度量文档间的相似性。文本处理中最常用的相似性度量方式是余弦距离。文本挖掘系统采用向量空间模型,用特征词条(T1 ,T2 ,…Tn) 及其权值Wi 代表目标信息,在进行信息匹配时,使用这些特征项评价未知文本与目标样本的相关程度。特征词条及其权值的选取称为目标样本的特征提取,特征提取算法的优劣将直接影响到系统的运行效果。

设D为一个包含m个文档的文档集合,Di为第i个文档的特征向量,则有

D={D1,D2,…,Dm},　Di=(di1,di2,…,din),i=1,2,…,m

其中dij(i=1,2,…,m;j=1,2,…,n)为文档Di中第j个词条tj的权值,它一般被定义为tj在Di中出现的频率tij的函数,例如采用TFIDF函数,即dij=tij\*log(N/nj)其中,N是文档[**数据库**](http://lib.csdn.net/base/mysql)中文档总数,nj是文档数据库含有词条tj的文档数目。假设用户给定的文档向量为Di,未知的文档向量为Dj,则两者的相似程度可用两向量的夹角余弦来度量,夹角越小说明相似度越高。相似度的计算公式如下:

通过上述的向量空间模型,文本数据就转换成了计算机可以处理的结构化数据,两个文档之间的相似性问题转变成了两个向量之间的相似性问题。

**三、 基于统计的特征提取方法(构造评估函数)**

**一、各种流行算法**

这类型算法通过构造评估函数,对特征集合中的每个特征进行评估,并对每个特征打分，这样每个词语都获得一个评估值,又称为权值。然后将所有特征按权值大小排序,提取预定数目的最优特征作为提取结果的特征子集。显然,对于这类型算法,决定文本特征提取效果的主要因素是评估函数的质量。

**1、TF-IDF：**

单词权重最为有效的实现方法就是TF\*IDF, 它是由Salton在1988 年提出的。其中TF 称为词频, 用于计算该词描述文档内容的能力; IDF 称为反文档频率, 用于计算该词区分文档的能力。TF\*IDF 的指导思想建立在这样一条基本假设之上: 在一个文本中出现很多次的单词, 在另一个同类文本中出现次数也会很多, 反之亦然。所以如果特征空间坐标系取TF 词频作为测度, 就可以体现同类文本的特点。另外还要考虑单词区别不同类别的能力, TF\*IDF 法认为一个单词出现的文本频率越小, 它区别不同类别的能力就越大, 所以引入了逆文本频度IDF 的概念, 以TF 和IDF 的乘积作为特征空间坐标系的取值测度。

TFIDF 法是以特征词在文档d中出现的次数与包含该特征词的文档数之比作为该词的权重，即

其中， Wi表示第i个特征词的权重，TFi(t，d)表示词t在文档d中的出现频率，N表示总的文档数，DF(t)表示包含t的文档数。用TFIDF算法来计算特征词的权重值是表示当一个词在这篇文档中出现的频率越高，同时在其他文档中出现的次数越少，则表明该词对于表示这篇文档的区分能力越强，所以其权重值就应该越大。将所有词的权值排序, 根据需要可以有两种选择方式:( 1) 选择权值最大的某一固定数n 个关键词;( 2) 选择权值大于某一阈值的关键词。一些实验表示,人工选择关键词, 4∽7 个比较合适, 机选关键词10∽15 通常具有最好的覆盖度和专指度。

TFIDF算法是建立在这样一个假设之上的：对区别文档最有意义的词语应该是那些在文档中出现频率高，而在整个文档集合的其他文档中出现频率少的词语，所以如果特征空间坐标系取TF词频作为测度，就可以体现同类文本的特点。另外考虑到单词区别不同类别的能力，TFIDF法认为一个单词出现的文本频数越小，它区别不同类别文本的能力就越大。因此引入了逆文本频度IDF的概念，以TF和IDF的乘积作为特征空间坐标系的取值测度，并用它完成对权值TF的调整，调整权值的目的在于突出重要单词，抑制次要单词。但是在本质上IDF是一种试图抑制噪音的加权 ，并且单纯地认为文本频数小的单词就越重要，文本频数大的单词就越无用，显然这并不是完全正确的。IDF的简单结构并不能有效地反映单词的重要程度和特征词的分布情况，使其无法很好地完成对权值调整的功能，所以TFIDF法的精度并不是很高。

此外，在TFIDF算法中并没有体现出单词的位置信息，对于Web文档而言，权重的计算方法应该体现出HTML的结构特征。特征词在不同的标记符中对文章内容的反映程度不同，其权重的计算方法也应不同。因此应该对于处于网页不同位置的特征词分别赋予不同的系数，然后乘以特征词的词频，以提高文本表示的效果。

**2、词频方法(Word Frequency)：**

词频是一个词在文档中出现的次数。通过词频进行特征选择就是将词频小于某一闭值的词删除，从而降低特征空间的维数。这个方法是基于这样一个假设，即出现频率小的词对过滤的影响也较小。但是在信息检索的研究中认为，有时频率小的词含有更多的信息。因此，在特征选择的过程中不宜简单地根据词频大幅度删词。

**3、文档频次方法(Document Frequency)：**

文档频数(Document Frequency, DF)是最为简单的一种特征选择算法,它指的是在整个数据集中有多少个文本包含这个单词。在训练文本集中对每个特征计一算它的文档频次，并且根据预先设定的阑值去除那些文档频次特别低和特别高的特征。文档频次通过在训练文档数量中计算线性近似复杂度来衡量巨大的文档集，计算复杂度较低，能够适用于任何语料，因此是特征降维的常用方法。

在训练文本集中对每个特征计算它的文档频数,若该项的DF 值小于某个阈值则将其删除,若其DF 值大于某个阈值也将其去掉。因为他们分别代表了“没有代表性”和“没有区分度”2 种极端的情况。DF 特征选取使稀有词要么不含有用信息,要么太少而不足以对分类产生影响,要么是噪音,所以可以删去。DF 的优点在于计算量很小,而在实际运用中却有很好的效果。缺点是稀有词可能在某一类文本中并不稀有,也可能包含着重要的判断信息,简单舍弃,可能影响分类器的精度。

文档频数最大的优势就是速度快,它的时间复杂度和文本数量成线性关系,所以非常适合于超大规模文本数据集的特征选择。不仅如此,文档频数还非常地高效,在有监督的特征选择应用中当删除90%单词的时候其性能与信息增益和x2 统计的性能还不相上下。DF 是最简单的特征项选取方法, 而且该方法的计算复杂度低, 能够胜任大规模的分类任务。

但如果某一稀有词条主要出现在某类训练集中,却能很好地反映类别的特征,而因低于某个设定的阈值而滤除掉,这样就会对分类精度有一定的影响。

**4、互信息(Mutual Information)：**

互信息衡量的是某个词和类别之间的统计独立关系,某个词t和某个类别Ci传统的互信息定义如下:

互信息是计算语言学模型分析的常用方法，它度量两个对象之间的相互性。在过滤问题中用于度量特征对于主题的区分度。互信息的定义与交叉嫡近似。互信息本来是信息论中的一个概念,用于表示信息之间的关系, 是两个随机变量统计相关性的测度，使用互信息理论进行特征抽取是基于如下假设:在某个特定类别出现频率高,但在其他类别出现频率比较低的词条与该类的互信息比较大。通常用互信息作为特征词和类别之问的测度，如果特征词属于该类的话，它们的互信息量最大。由于该方法不需要对特征词和类别之问关系的性质作任何假设，因此非常适合于文本分类的特征和类别的配准工作。

特征项和类别的互信息体现了特征项与类别的相关程度, 是一种广泛用于建立词关联统计模型的标准。互信息与期望交叉熵的不同在于没有考虑特征出现的频率, 这样导致互信息评估函数不选择高频的有用词而有可能选择稀有词作为文本的最佳特征。因为对于每一主题来讲，特征t的互信息越大，说明它与该主题的共现概率越大，因此，以互信息作为提取特征的评价时应选互信息最大的若干个特征。

互信息计算的时间复杂度类似于信息增益, 互信息的平均值就是信息增益。互信息的不足之处在于得分非常受词条边缘概率的影响。

实验数据显示,互信息分类效果最差,其次是文档频率、CC 统计,CHI 统计分类效果最好。

对互信息而言,提高分类精度的方法有:1) 可以增加特征空间的维数,以提取足够多的特征信息,这样就会带来了时间和空间上的额外开销;2) 根据互信息函数的定义,认为这些低频词携带着较为强烈的类别信息,从而对它们有不同程度的倚重. 当训练语料库没有达到一定规模的时候,特征空间中必然会存在大量的出现文档频率很低(比如低于3 次) 的词条,他们较低的文档频率导致了他们必然只属于少数类别. 但是从抽取出来的特征词观察发现,大多数为生僻词,很少一部分确实带有较强的类别信息,多数词携带少量的类别信息,甚至是噪音词.

**5、期望交叉熵(Expected Cross Entropy)：**

交叉嫡与信息量的定义近似，其公式为:

交叉嫡 ，也称KL距离。它反映了文本主题类的概率分布和在出现了某特定词汇的条件下文本主题类的概率分布之间的距离，词汇w的交叉嫡越大，对文本主题类分布的影响也越大。它与信息增益唯一的不同之处在于没有考虑单词未发生的情况，只计算出现在文本中的特征项。如果特征项和类别强相关, P ( Ci | w )就大,若P( Ci) 又很小的话,则说明该特征对分类的影响大。

交叉熵反映了文本类别的概率分布和在出现了某个特定词的条件下文本类别的概率分布之间的距离, 特征词t 的交叉熵越大, 对文本类别分布的影响也越大。熵的特征选择效果都要优于信息增益。

**6、二次信息熵(QEMI)：**

将二次熵函数应用于互信息评估方法中，取代互信息中的Shannon熵，就形成了基于二次熵的互信息评估函数。基于二次熵的互信息克服了互信息的随机性，是一个确定的量，因此可以作为信息的整体测度，另外它还比互信息最大化的计算复杂度要小，所以可以比较高效地用在基于分类的特征选取上。

二次熵的概念是在广义信息论中提出的。广义熵：

当 ，就得到了二次熵定义：

**7、信息增益方法(Information Gain)：**

信息增益方法是机器学习的常用方法，在过滤问题中用于度量已知一个特征是否出现于某主题相关文本中对于该主题预测有多少信息。通过计算信息增益可以得到那些在正例样本中出现频率高而在反例样本中出现频率低的特征，以及那些在反例样本中出现频率高而在正例样本中出现频率低的特征。信息增益G(w)的训算公式如下:

其中P(w)是词w出现的概率，P(Ci)是取第i个目录时的概率，P(C,|w ) 是假定w出现时取第i个目录的概率。

信息增益是一种基于熵的评估方法,涉及较多的数学理论和复杂的熵理论公式,定义为某特征项为整个分类所能提供的信息量,不考虑任何特征的熵与考虑该特征后的熵的差值。他根据训练数据,计算出各个特征项的信息增益,删除信息增益很小的项,其余的按照信息增益从大到小排序。

信息增益是信息论中的一个重要概念, 它表示了某一个特征项的存在与否对类别预测的影响, 定义为考虑某一特征项在文本中出现前后的信息熵之差。某个特征项的信息增益值越大, 贡献越大, 对分类也越重要。信息增益方法的不足之处在于它考虑了特征未发生的情况。特别是在类分布和特征值分布高度不平衡的情况下, 绝大多数类都是负类, 绝大多数特征都不出现。此时的函数值由不出现的特征决定, 因此, 信息增益的效果就会大大降低。信息增益表现出的分类性能偏低。因为信息增益考虑了文本特征未发生的情况，虽然特征不出现的情况肿可能对文本类别具有贡献，但这种贡献往往小于考虑这种情况时对特征分值带来的干扰。

**8、x2统计量方法：**

x2统计量用于度量特征w和主题类C之间的独立性。而表示除w以外的其他特征，C表示除C以外的其他主题类，那么特征w和主题类C的关系有以下四种

情况: ，用A, B, C, D表示这四种情况的文档频次，总的文档数N=A+B+C+D,扩统计量的计算公式如下:

当特征w和主题类C之间完全独立的时候，x2统计量为0。x2统计量和互信息的差别在于它是归一化的统计量，但是它对低频特征的区分效果也不好。X2 统计得分的计算有二次复杂度, 相似于互信息和信息增益。在 X2 统计和互信息之间主要的不同在于 X2 是规格化评价, 因而 X2 评估分值对在同类中的词是可比的, 但是 X2 统计对于低频词来说是不可靠的。

利用x2 统计方法来进行特征抽取是基于如下假设:在指定类别文本中出现频率高的词条与在其他类别文本中出现频率比较高的词条,对判定文档是否属于该类别都是很有帮助的.

采用x2估计特征选择算法的准确率在实验中最高，其分类效果受训练集影响较小，比较稳定。而且在对文教类和政治类存在类别交叉现象的文本进行分类时，采用x2估计的分类系统表现出了优于其它方法的分类性能。X2估计的可靠性较好，便于对程序的控制，无需因训练集的改变而人为的调节特征阀值的大小。

**9、文本证据权(The Weight of Evidence forText)：**

文本证据权衡量类的概率和给定特征时类的条件概率之间的差别。

**10、优势率(Odds Ratio)：**

    优势率只适用于二元分类的情况，其特点是只关心文本特征对于目标类的分值。Pos表示目标类，neg表示非目标类。

**11、遗传算法(Genetic Algorithm, GA)：**

文本实际上可以看作是由众多的特征词条构成的多维空间,而特征向量的选择就是多维空间中的寻优过程,因此在文本特征提取研究中可以使用高效寻优算法。遗传算法(Genetic Algorithm, GA)是一种通用型的优化搜索方法,它利用结构化的随机信息交换技术组合群体中各个结构中最好的生存因素,复制出最佳代码串,并使之一代一代地进化,最终获得满意的优化结果。在将文本特征提取问题转化为文本空间的寻优过程中,首先对Web文本空间进行遗传编码,以文本向量构成染色体,通过选择、交叉、变异等遗传操作,不断搜索问题域空间,使其不断得到进化,逐步得到Web文本的最优特征向量。

基于协同演化的遗传算法不是使用固定的环境来评价个体,而是使用其他的个体来评价特定个体。个体优劣的标准不是其生存环境以外的事物,而是由在同一生存竞争环境中的其他个体来决定。协同演化的思想非常适合处理同类文本的特征提取问题。由于同一类别文本相互之间存在一定相关性,因而各自所代表的那组个体在进化过程中存在着同类之间的相互评价和竞争。因此,每个文本的特征向量,即该问题中的个体,在不断的进化过程中,不仅受到其母体(文本)的评价和制约,而且还受到种族中其他同类个体的指导。所以,基于协同演化的遗传算法不仅能反映其母体的特征,还能反映其他同类文本的共性,这样可以有效地解决同一主题众多文本的集体特征向量的提取问题,获得反映整个文本集合某些特征的最佳个体。

**12、主成分分析法(Principal Component Analysis，PCA)：**

它不是通过特征选取的方式降维的，而是通过搜索最能代表原数据的正交向量，创立一个替换的、较小的变量集来组合属性的精华，原数据可以投影到这个较小的集合。PCA由于其处理方式的不同又分为数据方法和矩阵方法。矩阵方法中，所有的数据通过计算方差一协方差结构在矩阵中表示出来，矩阵的实现目标是确定协方差矩阵的特征向量，它们和原始数据的主要成分相对应。在主成分方法中，由于矩阵方法的复杂度在n很大的情况 以二次方增长，因此人们又开发使用了主要使用Hebbian学习规则的PCA神经网络方法。

主成分分析法是特征选取常用的方法之一，它能够揭示更多有关变量\_丰要方向的信息。但它的问题在于矩阵方法中要使用奇异值分解对角化矩阵求解方差一协方差。

**13、模拟退火算法(Simulating Anneal，SA)：**

特征选取可以看成是一个组合优化问题，因而可以使用解决优化问题的方法来解决特征选取的问题。模拟退火算法(Simulating Anneal，SA)就是其中一种方法。

模拟退火算法是一个很好的解决优化问题的方法，将这个方法运用到特征选取中，理论上能够找到全局最优解，但在初始温度的选取和邻域的选取t要恰当，必须要找到一个比较折中的办法，综合考虑解的性能和算法的速度。

**14、N—Gram算法**

它的基本思想是将文本内容按字节流进行大小为N的滑动窗口操作,形成长度为N的字节片段序列。每个字节片段称为gram,对全部gram的出现频度进行统计,并按照事先设定的阈值进行过滤,形成关键gram列表,即为该文本的特征向量空间,每一种gram则为特征向量维度。由于N—Gram算法可以避免汉语分词的障碍,所以在中文文本处理中具有较高的实用性。中文文本处理大多采用双字节进行分解,称之为bi-gram。但是bigram切分方法在处理20%左右的中文多字词时,往往产生语义和语序方面的偏差。而对于专业研究领域,多字词常常是文本的核心特征,处理错误会导致较大的负面影响。基于N—Gram改进的文本特征提取算法[2],在进行bigram切分时,不仅统计gram的出现频度,而且还统计某个gram与其前邻gram的情况,并将其记录在gram关联矩阵中。对于那些连续出现频率大于事先设定阈值的,就将其合并成为多字特征词。这样通过统计与合并双字特征词,自动产生多字特征词,可以较好地弥补N—Gram算法在处理多字词方面的缺陷。

**15、各种方法的综合评价**

上述几种评价函数都是试图通过概率找出特征与主题类之间的联系，信息增益的定义过于复杂，因此应用较多的是交叉嫡和互信息。其中互信息的效果要好于交又嫡，这是因为互信息是对不同的主题类分别抽取特征词，而交叉嫡跟特征在全部主题类内的分布有关，是对全部主题类来抽取特征词。这些方法,在英文特征提取方面都有各自的优势,但用于中文文本,并没有很高的效率。主要有2 个方面的原因:1) 特征提取的计算量太大,特征提取效率太低,而特征提取的效率直接影响到整个文本分类系统的效率。2) 经过特征提取后生成的特征向量维数太高,而且不能直接计算出特征向量中各个特征词的权重。

目前使用评估函数进行特征选取越来越普遍，特征选取算法通过构造一个评估函数的方法，选取预定数目的最佳特征作为特征子集的结果。在几种评估方法中，每一种方法都有一个选词标准，遵从这个标准，从文本集的所有词汇中选取出有某个限定范围的特征词集。因为评估函数的构造不是特别复杂，适用范围又很广泛，所以越来越多的人们喜欢使用构造评估函数来进行特征的选取。

这些评估函数在Web文本挖掘中被广泛使用,特征选择精度普遍达到70%~80%,但也各自存在缺点和不足。例如,“信息增益”考虑了单词未发生的情况,对判断文本类别贡献不大,而且引入不必要的干扰,特别是在处理类分布和特征值分布高度不平衡的数据时选择精度下降。“期望交叉熵”与“信息增益”的唯一不同就是没有考虑单词未发生的情况,因此不论处理哪种数据集,它的特征选择精度都优于“信息增益”。与“期望交叉熵”相比,“互信息”没有考虑单词发生的频度,这是一个很大的缺点,造成“互信息”评估函数经常倾向于选择稀有单词。“文本证据权”是一种构造比较新颖的评估函数,它衡量一般类的概率和给定特征类的条件概率之间的差别,这样在文本处理中,就不需要计算W的所有可能值,而仅考虑W在文本中出现的情况。“优势率”不像前面所述的其他评估函数将所有类同等对待,它只关心目标类值,所以特别适用于二元分类器,可以尽可能多地识别正类,而不关心识别出负类。

从考虑文本类间相关性的角度,可以把常用的评估函数分为两类,即类间不相关的和类间相关的。“文档频数”(DF)是典型的类间不相关评估函数, DF的排序标准是依据特征词在文档中出现篇数的百分比,或称为篇章覆盖率。这种类型的评估函数,为了提高区分度,要尽量寻找篇章覆盖率较高的特征词,但又要避免选择在各类文本中都多次出现的无意义高频词,因此类间不相关评估函数对停用词表的要求很高。但是,很难建立适用于多个类的停用词表,停用词不能选择太多,也不能选择太少,否则都将会影响特征词的选择。同时,类间不相关评估函数还存在一个明显的缺点,就是对于特征词有交叉的类别或特征相近的类别,选择的特征词会出现很多相似或相同的词条,造成在特定类别间的区分度下降。类间相关的评估函数,例如期望交叉熵、互信息、文本证据权等,综合考虑了词条在已定义的所有类别中的出现情况,可以通过调整特征词的权重,选择出区分度更好的特征,在一定程度上提高了相近类别的区分度。但是，该区分度的提高仅体现在已定义的类别间,而对于尚未定义的域外类别,类间相关评估函数的选择效果也不理想。因此,在评估函数选择问题上,提高对域外类别文本的区分度是十分重要的研究课题。

传统的特征选择方法大多采用以上各评估函数进行特征权重的计算，由于这些评估函数是基于统计学的，其中一个主要缺陷就是需要用一个很庞大的训练集才能获得几乎所有的对分类起关键作用的特征．这需要消耗大量的时间和空间资源，况且，构建这样一个庞大的训练集也是一项十分艰巨的工作。然而，在现实应用中，考虑到工作效率，不会也没有足够的资源去构建一个庞大的训练集，这样的结果就是：被选中的甚至是权重比较高的特征，可能对分类没有什么用处，反而会干涉到正确的分类；而真正有用的特征却因为出现的频率低而获得 较低的权重，甚至在降低特征空间维数的时候被删除掉了。

基于评估函数的特征提取方法是建立在特征独立的假设基础上,但在实际中这个假设是很难成立的,因此需要考虑特征相关条件下的文本特征提取方法。

**二、影响特征词权值的因素分析**

**1、词频**

    文本内空中的中频词往往具有代表性，高频词区分能力较小，而低频词或者示出现词也常常可以做为关键特征词。所以词频是特征提取中必须考虑的重要因素，并且在不同方法中有不同的应用公式。

**2、词性**

汉语言中，能标识文本特性的往往是文本中的实词，如名词、动词、形容词等。而文本中的一些虚词，如感叹词、介词、连词等，对于标识文本的类别特性并没有贡献，也就是对确定文本类别没有意义的词。如果把这些对文本分类没有意思的虚词作为文本特征词，将会带来很大噪音，从而直接降低文本分类的效率和准确率。因此，在提取文本特征时，应首先考虑剔除这些对文本分类没有用处的虚词，而在实词中，又以名词和动词对于文本的类别特性的表现力最强，所以可以只提取文本中的名词和动词作为文本的一级特征词。

**2、文档频次**

    出现文档多的特征词，分类区分能力较差，出现文档少的特征词更能代表文本的不同主题。

**2、标题**

标题是作者给出的提示文章内容的短语，特别在新闻领域，新闻报道的标题一般都要求要简练、醒目，有不少缩略语，与报道的主要内容有着重要的联系，对摘要内容的影响不可忽视。统计分析表明,小标题的识别有助于准确地把握文章的主题。主要体现在两个方面:正确识别小标题可以很好地把握文章的整体框架,理清文章的结构层次;同时,小标题本身是文章中心内容的高度概括。因此,小标题的正确识别能在一定程度上提高文摘的质量。

**3、位置**

美国的EE．Baxendale的调查结果显示：段落的论题是段落首句的概率为85％ ， 是段落末句的概率为7％ 。而且新闻报道性文章的形式特征决定了第一段一般是揭示文章主要内容的。因此，有必要提高处于特殊位置的句子权重，特别是报道的首旬和末句。但是这种现象又不是绝对的，所以，我们不能认为首句和末句就一定是所要摘要的内容，因此可以考虑一个折衷的办法，即首句和末句的权重上可通过统计数字扩大一个常数倍。首段、末段、段首、段尾、标题和副标题、子标题等处的句子往往在较大程度上概述了文章的内容。对于出现在这些位置的句子应该加大权重。

Internet上的文本信息大多是HTML结构的，对于处于Web文本结构中不同位置的单词，其相应的表示文本内容或区别文本类别的能力是不同的，所以在单词权值中应该体现出该词的位置信息。

**4、句法结构**

句式与句子的重要性之间存在着某种联系，比如摘要中的句子大多是陈述句，而疑问句、感叹句等则不具内容代表性。而通常“总之”、“综上所述”等一些概括性语义后的句子，包含了文本的中心内容。

**5、专业词库**

通用词库包含了大量不会成为特征项的常用词汇,为了提高系统运行效率,系统根据挖掘目标建立专业的分词表,这样可以在保证特征提取准确性的前提下,显著提高系统的运行效率。

用户并不在乎具体的哪一个词出现得多,而在乎泛化的哪一类词出现得多。真正起决定作用的是某一类词出现的总频率。基于这一原理,我们可以先将词通过一些方法依主题领域划分为多个类,然后为文本提取各个词类的词频特征,以完成对文本的分类。

可以通过人工确定领域内的关键词集。

**6、信息熵**

熵(Entropy)在信息论中是一个非常重要的概念 ' ，它是不确定性的一种度量。信息熵方法的基本目的是找出某种符号系统的信息量和多余度之间的关系，以便能用最小的成本和消耗来实现最高效率的数据储存、管理和传递。信息熵是数学方法和语言文字学的结合，其定义为：设x是取有限个值的随机变量，各个取值出现的概率为 则 的熵为 其中，底数n可以为任意正数，并规定当 时， =0。在式(3)中，对数底a决定了熵的单位，如a=2、e、10，熵的单位分别为Bit，nat，Hartley。在我们的研究论文中，均取a＝2。熵具有最大值和最小值 ，由熵的定义公式可以看出，当每个值出现的概率相等时，即当 时这时熵函数达到最大值 ，记为最大熵 。其中Pt ≥0，并且 。而当 n)时，熵值最小，Entropy(X)=0。

我们将可以将信息论中的熵原理引入到特征词权重的计算中。

**7、文档、词语长度**

一般情况下,词的长度越短,其语义越泛。一般来说, 中文中词长较长的词往往反映比较具体、下位的概念, 而短的词常常表示相对抽象、上位的概念一般说来, 短词具有较高的频率和更多的含义, 是面向功能的;而长词的频率较低, 是面向内容的, 增加长词的权重, 有利于词汇进行分割, 从而更准确地反映出特征词在文章中的重要程度。词语长度通常不被研究者重视。但是本文在实际应用中发现,关键词通常是一些专业学术组合词汇,长度较一般词汇长。考虑候选词的长度,会突出长词的作用。长度项也可以使用对数函数来平滑词汇间长度的剧烈差异。通常来说,长词汇含义更明确,更能反映文本主题,适合作为关键词,因此将包含在长词汇中低于一定过滤阈值的短词汇进行了过滤。所谓过滤阈值,就是指进行过滤短词汇的后处理时,短词汇的权重和长词汇的权重的比的最大值。如果低于过滤阈值,则过滤短词汇,否则保留短词汇。

根据统计,二字词汇多是常用词,不适合作为关键词,因此对实际得到的二字关键词可以做出限制。比如,抽取5 个关键词,本文最多允许3 个二字关键词存在。这样的后处理无疑会降低关键词抽取的准确度和召回率,但是同候选词长度项的运用一样,人工评价效果将会提高。

**8、词语间关联**

**9、单词的区分能力**

在TF\*IDF 公式的基础上, 又扩展了一项单词的类区分能力。新扩展的项用于描述单词与各个类别之间的相关程度。

**10、词语直径(Diameter ( t) )**

词语直径是指词语在文本中首次出现的位置和末次出现的位置之间的距离。词语直径是根据实践提出的一种统计特征。根据经验,如果某个词汇在文本开头处提到,结尾又提到,那么它对该文本来说,是个很重要的词汇。不过统计结果显示,关键词的直径分布出现了两极分化的趋势,在文本中仅仅出现了1 次的关键词占全部关键词的14.184 %。所以词语直径是比较粗糙的度量特征。

**11、首次出现位置(FirstLoc ( t) )**

Frank 在Kea 算法中使用候选词首次出现位置作为Bayes 概率计算的一个主要特征,他称之为距离(Distance)。简单的统计可以发现,关键词一般在文章中较早出现,因此出现位置靠前的候选词应该加大权重。实验数据表明,首次出现位置和词语直径两个特征只选择一个使用就可以了。由于文献数据加工问题导致中国学术期刊全文数据库的全文数据不仅包含文章本身,还包含了作者、作者机构以及引文信息,针对这个特点,使用首次出现位置这个特征,可以尽可能减少全文数据的附加信息造成的不良影响。

**12、词语分布偏差(Deviation ( t) )**

词语分布偏差所考虑的是词语在文章中的统计分布。在整篇文章中分布均匀的词语通常是重要的词汇。词语的分布偏差计算公式如下: 其中,CurLoc ( tj ) 是词汇t 在文章中第j 次出现的位置;MeanLoc ( t ) 是词汇t 在文章中出现的平均位置。

**13、特征提取的一般步骤**

一、 候选词的确定

 (1) 分词(词库的扩充)

尽管现在分词软件的准确率已经比较高了，但是，它对专业术语的识别率还是很好，所以，为了进一步提高关键词抽取的准确率，我们有必要在词库中添加了一个专业词库以保证分词的质量。

(2) 停用词的过滤

停用词是指那些不能反映主题的功能词。例如：“的”、“地”、“得”之类的助词，以及像“然而”、“因此”等只能反映句子语法结构的词语，它们不但不能反映文献的主题，而且还会对关键词的抽取造成干扰，有必要将其滤除。停用词确定为所有虚词以及标点符号。

(3) 记录候选词在文献中的位置

为了获取每个词的位置信息，需要确定记录位置信息的方式以及各个位置的词在反映主题时的相对重要性。根据以往的研究结果，初步设定标题的位置权重为5，摘要和结论部分为3，正文为1，同时，把标题、摘要和结论、正文分别称为5 区、3 区和1 区。确定了文章各个部分的位置权重之后，就可以用数字标签对每个位置做一个标记。做法是：在标题的开头标上数字5，在摘要和结论部分的段首标上数字3，在正文的段首标上数字1，这样，当软件逐词扫描统计词频时，就可以记录每个词的位置信息。

二、词语权重计算

(1) 词语权值函数的构造(见各不同算法)

(2) 关键词抽取

候选词的权值确定以后，将权值排序，取前n个词作为最后的抽取结果。

**四、基于语义的特征提取方法（结合领域）**

**一、基于语境框架的文本特征提取方法**

越来越多的现象表明，统计并不能完全取代语义分析。不考虑句子的含义和句子间的关系机械抽取，必然导致主题的准确率低，连贯性差，产生一系列问题，如主要内容缺失、指代词悬挂、文摘句过长等。因此，理想的自动主题提取模型应当将两种方法相结合。应当将语义分析融入统计算法，基本的方法仍然是“统计-抽取”模型，因为这一技术已经相对成熟并拥有丰富的研究成果。

语境框架是一个三维的语义描述,把文本内容抽象为领域(静态范畴)、情景(动态描述)、背景(褒贬、参照等)三个框架。在语境框架的基础上,从语义分析入手,实现了4元组表示的领域提取算法、以领域句类为核心的情景提取算法和以对象语义立场网络图为基础的褒贬判断。该方法可以有效地处理语言中的褒贬倾向、同义、多义等现象,表现出较好的特征提取能力。

**二、基于本体论的文本提取方法**

应用本体论(On-tology)模型可以有效地解决特定领域知识的描述问题。具体针对数字图像领域的文本特征提取,通过构建文本结构树,给出特征权值的计算公式。算法充分考虑特征词的位置以及相互之间关系的分析,利用特征词统领长度的概念和计算方法,能够更准确地进行特征词权值的计算和文本特征的提取。

**三、基于知网的概念特征提取方法**

对于文本的处理,尤其是中文文本处理,字、词、短语等特征项是处理的主要对象。但是字、词、短语更多体现的是文档的词汇信息,而不是它的语义信息,因而无法准确表达文档的内容;大多数关于文本特征提取的研究方法只偏重考虑特征发生的概率和所处的位置,而缺乏语义方面的分析;向量空间模型最基本的假设是各个分量间正交,但作为分量的词汇间存在很大的相关性,无法满足模型的假设。基于概念特征的特征提取方法是在VSM的基础上,对文本进行部分语义分析,利用知网获取词汇的语义信息,将语义相同的词汇映射到同一概念,进行概念聚类,并将概念相同的词合并成同一词。用聚类得到的词作为文档向量的特征项,能够比普通词汇更加准确地表达文档内容,减少特征之间的相关性和同义现象。这样可以有效降低文档向量的维数,减少文档处理计算量,提高特征提取的精度和效率。

**五、可创新点**

一、文本特征提取及文本挖掘在军事情报领域的应用研究

二、新的文本特征表示模型

考虑使用二级向量进行文本特征建模,关键词向量能快速定位用户的兴趣领域,而扩展词向量能准确反映用户在该领域上的兴趣偏好。结合领域知识，采用概念词、同义词或本体来代替具体的关键词成为特征词，体现语义层面的需求和分析。

三、军事情报领域专业主题词库、敏感词库的构造

    通过人机结合的方式，建立军事情报领导的最佳关键词表，即主题词库，或者语义库。将专业词库应用于军事系统的分词、特征提取、分类、信息抽取和监控、文本挖掘等方面。

四、改进分词算法

    利用专业词库改进中科院ICTCLAS分词系统([**Java**](http://lib.csdn.net/base/java)开源版)，并针对特征提取需要，改进分词步骤，直接在分词过程中剔除无意义词语，增加词语附加信息，集成相关因素影响值的计算，加快系统速度和效率。

五、改进特征评价函数

    将影响因素中的经验知识通过实验和验算定量表示(如位置影响、高频词权值增加)，形成可行的非性线加权公式。探索二次迭代算法，在初次文本分析基础上，记录文本的结构组成特征，记录候选特征词，记录文档集的组成特征，然后根据加权理论调整算法，第二次计算词的权值，最终提取高准确度的特征词集。