中文文本处理算法介绍

# 引言

本系统是对校园新闻为主体的中文文本处理，处理方法包括文本自动分类、文章标签提取、文章自动摘要、为用户智能推荐。本系统通过校园数据的再整理，方便用户能在浏览新闻时快速找到或者发现自己可能感兴趣的内容，大大节约用户获取信息的时间，提高用户的阅读体验。本系统同样适用于其他主题的中文文本领域，可以简单的移植到其他领域。

本系统建立中文文本数据数学描述的过程分为三个步骤：文本预处理、建立向量空间模型和空间向量模型的优化。文本预处理主要采用分词、停用词过滤等技术将原始的文本字符串转化为词条串。文本预处理之后，每一个文本的词条串被进一步转换为一个文本向量，也就是向量空间模型。向量的每一维对应一个词条，其值反映的是这个词条与这个文本之间的相似度。相似度有很多不同的计算方法，所以优化文本向量就是采用最为合适的计算方法来规范化文本向量，使其能更好地应用于文本分类和文本聚类等方面，本系统使用了TF-IDF算法对空间向量进行优化和计算词的权重。

在建立了中文文本数据数学描述后，就可以提取文本信息了，本系统使用的文本处理方法有文本自动摘要、文本关键字提取、文本自动摘要、用户阅读推荐。

# 中文文本分词

## 分词词典

分词算法中带有一个叫做dict.txt的词典，里面有2万多条词，包含了词条出现的次数(这个次数是基于人民日报语料等资源训练得出来的)和词性。把这2万多条词使用trie这种数据结构存储，trie树，可以叫前缀树，有时也称字典树，是字符串算法中比较常用的一种结构。如果使用 trie 树来实现英文单词的查找，那么最终形成的结构，如下图所示：

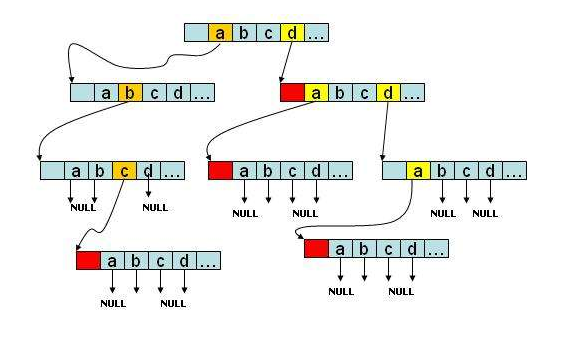


图 1 trie树查询结构

也就是说一个词语的前面几个字一样, 就表示他们具有相同的前缀, 就可以使用trie树来存储, 具有查找速度快的优势。

## 形成切分词图

对待分词句子，为了消除分词中的歧异，提高切分准确度，需要找出一句话所有可能的分词结果，也就是根据dict.txt生成的trie树，生成全切分词图。

如果待切分的字符串有m个字符，考虑每个字符左边和右边的位置，则有m+1个点对应，点的编号从0到m。把候选词看成边，可以根据词典生成一个切分词图。切分词图是一个有向正权重的图。例如"有意见分歧"这句话的切分词图如图所示。

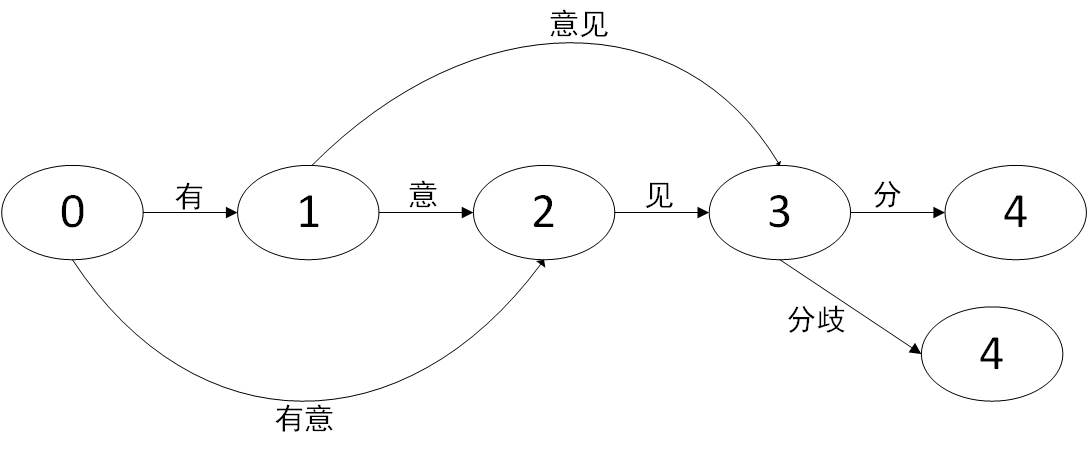


图 2 全切分词图

在"有意见分歧"的切分词图中："有"这条边的起点是0，终点是1；"有意"这条边的起点是0，终点是2，以此类推。切分方案就是从源点0到终点5之间的路径，共存在两条切分路径。

路径1：0－1－3－5，对应切分方案S1：有/意见/分歧/。

路径2：0－2－3－5，对应切分方案S2：有意/见/分歧/。

## 查找最大概率路径

在2.2中介绍对于一个特定的字符串C，会有多个切分方案S对应，分词的任务就是在这些S中找出概率最大的一个切分方案，也就是对输入字符串切分出最有可能的词序列。



计算条件概率和，然后采用概率大的值对应的切分方案。根据贝叶斯公式，有

 。

其中是字符串在语料库中出现的概率，只是一个用来归一化的固定值。从词串恢复到汉字串的概率只有唯一的一种方式，所以。因此，比较和的大小变成比较和的大小。

概率语言模型分词的任务是：在全切分所得的所有结果中求某个切分方案S，使得P(S)最大。那么，如何来表示P(S)呢？为了容易实现，假设每个词之间的概率是上下文无关的，则：

其中，对于不同的S，m的值是不一样的，一般来说m越大，P(S)会越小。也就是说，分出的词越多，概率越小。这符合实际的观察，如最大长度匹配切分往往会使得m较小。计算任意一个词出现的概率如下：



因此



从另外一个角度来看，计算最大概率等于求切分词图的最短路径。

常见的词语概率表如表所示：

|  |  |
| --- | --- |
| 词语 | 概率 |
| … | … |
| 有 | 0.0180 |
| 有意 | 0.0005 |
| 意见 | 0.0010 |
| 见 | 0.0002 |
| 分歧 | 0.0001 |
| … | … |





可得，所以选择对应的切分。

## 算法优化

如何尽快找到概率最大的词串？因为假设每个词之间的概率是上下文无关的，因此满足用动态规划求解所要求的最优子结构性质和无后效性，可以采用动态规划的方法求解最短路径。在动态规划求解的过程中并没有先生成所有可能的切分路径，而是求出值最大的后，利用回溯的方法直接输出。

到节点为止的最大概率称为节点的概率：



如果的结束节点是，就称为的前驱词。这里的就是节点的前驱词集合。比如上面的例子中，候选词"有"就是节点1的前驱词，"意见"和"见"都是节点3的前驱词。是的开始节点，也是节点的前驱节点。因此切分的最大概率就是(节点m的最佳前驱节点) P(节点m的最佳前驱词)。

# 建立空间向量模型

计算机并不具有人类的智能，人在阅读文章后, 根据自身的理解能力可以产生对文章内容的模糊认识；而计算机并不能轻易地“读懂”文章，从根本上说，它只认识0和1，所以必须将文本转换为计算机可以识别的格式。根据“贝叶斯假设”，假定组成文本的字或词在确定文本类别的作用上相互独立，这样，可以就使用文本中出现的字或词的集合来代替文本。不言而喻, 这将丢失大量关于文章内容的信息，但是这种假设可以使文本的表示和处理形式化，并且可以在文本分类中取得较好的效果。

目前，在信息处理方向上,文本的表示主要采用向量空间模型(VSM)。向量空间模型的基本思想是以向量来表示文本:(W1, W2, W3, …,Wn),其中Wi为第i个特征项的权重。那么选取什么作为特征项呢? 一般可以选择字、词或词组。根据实验结果, 普遍认为选取词作为特征项要优于字和词组, 因此,要将文本表示为向量空间中的一个向量,就首先要将文本分词, 由这些词作为向量的维数来表示文本。

# 基于TF-IDF算法的空间向量模型优化

## TF-IDF算法介绍

TF-IDF（Term Frequency-Invers Document Frequency）算法是一种常用于信息处理和数据挖掘的加权算法。该算法采用一种统计方法，根据词条在文本中出现的次数和在整个语料中出现的文档频率来计算一个词条在整个语料中的重要程度。它的优点是能过滤掉一些常见的却无关紧要本的词语，同时保留影响整个文本的重要字词。

在一份给定的文件里，词频（term frequency，TF）指的是某一个给定的词语在该文件中出现的频率。这个数字是对词数(term count)的归一化，以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词数，而不管该词语重要与否。）对于在某一特定文件里的词语来说，它的重要性可表示为：



以上式子中  是该词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和。

逆向文件频率（inverse document frequency，IDF）是一个词语普遍重要性的度量。某一特定词语的，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到：



其中，表示语料库中的文件总数，表示词条的文件数目，如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用。然后



某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语。

## 模型优化

### 词条权值优化

最初的空间向量模型中向量的表示完全是0 , 1形式，即如果文本中出现了该词，那么文本向量的该维为1，否则为0。这种方法无法体现这个词在文本中的作用程度，所以0, 1逐渐被更精确的词频代替，词频分为绝对词频和相对词频。绝对词频，即使用词在文本中出现的频率表示文本，但是词频表示一个词在作用程度是不准确的，比如“你”、“我”、“他”这类词在所有文章中出现的频率都很高，但是不能特定的表示某类文章。相对词频，即使用某种统计算法得到词的相对作用程度。本项目运用TF-IDF算法计算某个词在文章中的重要程度，利用TF-IDF算法的优点就可以准确的得到文章中所有词的作用程度。

### 空间向量降维

在空间向量模型中，为了在同一空间向量中比较所有的文章，需要建立整个语料库的空间向量模型，也就是空间向量的维度是整个语料库中出现的词数。常见的中文词即使去掉停用词也超过2万条，也就意味着空间向量的维度大于2万，现在的计算资源很难处理这么大的数据量，高维的空间向量导致了维度灾难的出现。

在TF-IDF算法中需要计算词的逆向文件频率（inverse document frequency，IDF），也就是某个词在全部文章中出现的频率，如果一个词的逆向文件频率很大，说明这个词在很多文章中都出现过，意味着这个词对特定文章的表示能力很弱。本项目设定了特定的阈值，将逆向文件频率很高的词在空间向量中移除。这样就可以在最大程度保留文本特征的前提下大大降低空间向量模型的维度，大大优化了模型。

# 分类算法

分类是数据挖掘中应用领域极其广泛的重要技术之一，至今已经提出很多算法。分类是根据数据集的特点构造一个分类器，利用分类器对未知类别的样本赋予类别的一种技术。构造分类器的过程一般分为训练和测试两个步骤。在训练阶段，分析训练数据集的特点，为每个类别产生一个对相应数据集的准确描述或模型。在测试阶段，利用类别的描述或模型对测试进行分类，测试其分类准确度。一般来说,测试阶段的代价远远低于训练阶段。

本项目所需要的分类器是针对校园信息的文本分类，类别限定在学科竞赛、学术信息、工大新闻、招聘就业、校园活动、通知公告、考研留学、生活娱乐这八类，足以归纳所有的校园信息。目的是做一个程序，输入一篇文章的内容，输出这篇文章的类型。

本项目基于上述建立的空间向量模型，结合特定的分类算法构建分类器就对文章进行分类。分类算法流程为收集数据、建立空间向量模型、构建分类器。空间向量模型的已在上文中详细介绍，不在本节重复介绍。

## 收集数据

收集数据包括训练数据和测试数据的收集，测试数据的收集成本较训练数据的收集成本低很多。

训练数据是一定数量的已知类型的文本数据，由两部分组成：数据和标签，在这里数据就是文章内容，标签就是文章类型。训练数据作为所有文章特征的体现者，数量如果太少，不足以反映出所有待分类数据的特征，如果太多，可能出现过拟合现象。而且每个类型的训练数据量应该是相近的。最重要的是，训练数据的固有类型和标签一定要真实对应的，否则训练好的分类器不符合自然逻辑。

网上没有现成的针对校园信息的数据集，所以训练数据和测试数据的收集是项目组成员手动进行的，项目组成员依次审核分类结果是否符合自然逻辑，在最大程度上保证了文本内容与标签的一致性。最后每类文章标注了各100条校园新闻一共16000条数据分别作为训练数据和测试数据。

## 基于朴素贝叶斯的分类器构建

### 贝叶斯理论

贝叶斯决策理论方法是统计模型决策中的一个基本方法，公式为：



**、**分别表示事件A和事件B发生的概率，表示事件B已经发生的前提下，事件A发生的概率，叫做事件B发生下事件A的条件概率，同理，表示事件A已经发生的前提下，事件B发生的概率，叫做事件A发生下事件B的条件概率。

贝叶斯定理之所以有用，是因为我们在生活中经常遇到这种情况：我们可以很容易直接得出，则很难直接得出，但我们更关心，贝叶斯定理就为我们打通从P(A|B)获得P(B|A)的道路。

### 朴素贝叶斯分类

算法流程：

1. 设为一个待分类项，而每个a为x的一个特征属性。
2. 有类别集合。
3. 计算,,…,。
4. 找到一个已知分类的训练样本集。
5. 统计得到在各类别下各个特征属性的条件概率估计。即。

条件概率表示在类别为y时特征属性为a的概率。当特征属性为离散值时，可以通过统计训练样本中各个划分在每个类别中出现的频率来估。当特征属性为连续值时，通常假定其值服从高斯分布。即：





其中，、分别为训练样本中各个类别中此特征项的均值和标准差。使用上述公式即可得到需要的条件概率。

1. 在4）中可能出现的情况，当某个类别下某个特征项划分没有出现时，就是产生这种现象，这会令分类器质量大大降低。为了解决这个题，我们引入Laplace校准，它的思想非常简单，就是对没类别下所有划分的计数加1，这样如果训练样本集数量充分大时，并不会对结果产生影响，并且解决了上述频率为0使分类质量降低的问题。
2. 根据贝叶斯定理有如下推导：



因为各特征属性是条件独立的，所以有：



通过这步骤1）2）3）4）后，就可以计算得到某个样本属于每一类的概率。

1. 如果，则。

## 分类结果评估

根据对测试数据的测试，分类器的效果很准确，但对某些类别不明确的文本效果较差，在可以接受的范围内。通过实际测试，分类的正确率可达到96.3%，可以达到使用的要求。

# 关键字提取

关键字提取是在一篇文档中找到最能体现这篇文档几个关键字，可以帮助读者判断是否对这篇文章感兴趣，也可以用作关键词索引。

本项目基于上述建立的空间向量模型，利用TF-IDF算法得到文本所有词条的权值，统计权值前五的5个词作为文章的关键词。

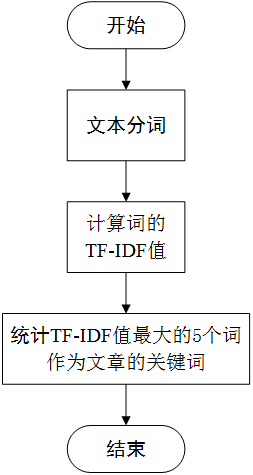


图 3文章关键词提取算法流程图

# 自动摘要

文章自动摘要是提取一篇文档中的摘要，可以帮助读者快速浏览这篇文章的主要内容。本项目使用了基于词频统计的文章自动摘要。原理是文章是由句子组成的，文章的信息都包含在句子中，有些句子包含的信息多，有些句子包含的信息少。句子的信息量用"关键词"来衡量。如果包含的关键词越多，就说明这个句子越重要。自动摘要就是要找出那些包含信息最多的句子，也就是包含关键字最多的句子。而通过统计句子中关键字的频率的大小，进而进行排序，通过对排序的词频列表对文档中句子逐个进行打分，进而把打分高的句子找出来，就可以作为这篇文章的摘要。

# 个性化推荐

个性化是通过挖掘用户的历史记录和浏览记录，将这些记录与用户信息相结合，推荐出用户可能感兴趣的事物。这样既给用户节约时间搜索信息，又让用户体验到智慧推荐方便快捷的优点。

目前主要的推荐技术分为，协同过滤推荐，对用户做出相似比较，为相似度高的用户推荐相同的内容，使用前提是积累一定量的用户浏览信息；基于内容推荐，在用户浏览操作记录的基础上推荐用户可能感兴趣的内容。本项目将协同过滤和基于内容过滤两种算法取长补短进行组合，设计出混合推荐算法。混合推荐算法的主要过程如下：

在用户浏览信息积累很少时，使用基于内容的推荐算法，本项目使用文章的关键字表征文章内容，为用户推荐有相同关键字的文章。

在积累足量的用户浏览信息后，使用基于协同过滤推荐算法推荐用户可能感兴趣的文章，本项目采用余弦相似度算法计算用户相似性。

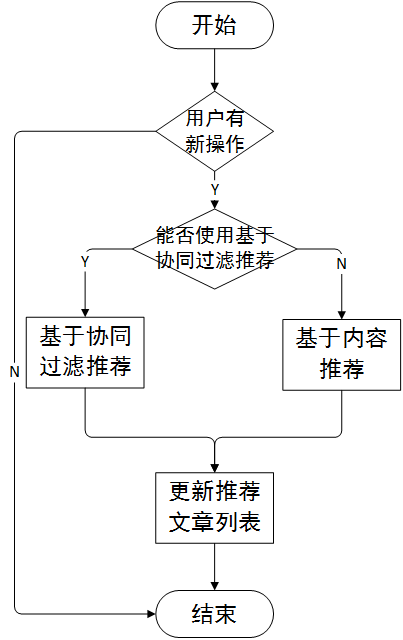


图 4智能推荐算法流程图

个性化推荐将自动为用户筛选感兴趣的话题，例如为热衷科研竞赛的用户推送竞赛的详细信息，为毕业生提供实习招聘信息。个性化推荐是综合推荐用户可能感兴趣的各类信息，并将这些信息发布到同一个界面中，用户不需要跳转界面就可以浏览自己需要和感兴趣的各类信息，可以是用户短时间内获得自己感兴趣的所有类型文章。