# 文本分类技术及在军事情报中的应用

赵国际[1],李竹林[2],赵宗涛[2],张 宏[3]
(1.西北大学 信息学院,陕西 西安 710069;
2.第二炮兵工程学院,陕西 西安 710025;
3.延安市教研室,陕西 延安 716000)

**摘 要**:作战数据库文件与记录的文本格式涉及到作战文书的自动生成,文本分类直接关系到情报信息检索效率及准确性。针对军事情报信息的特点,建立了一个基于情报数据库的文本分类模型,然后分析了模型中的文本表示、自动分词、特征提取关键技术,并对互信息特征选取方法提出了改进措施。应用表明,该文本分类模型可有效地从文字信息中分离出规范化的情报要点,不仅辅助作战决策,而且能直接写入数据库。

**关键词**:情报管理;文本分类;互信息
**中图分类号**:TP391 **文献标识码**:A **文章编号**:1673-629X(2007)08-0176-04

# Technology of Text Classification and Application in Military Intelligence Management

ZHAO Guo-ji[1], LI Zhu-lin[2], ZHAO Zong-tao[2], ZHANG Hong[3]
(1. Department of Information, Northwest University, Xi'an 710069, China;
2. The Second Artillery Engineering College, Xi'an 710025, China;
3. Teaching and Research Division of Yan'an, Yan'an 716000, China)

**Abstract**: Files and records of text format for battle database are important, it relates to automatic generating of operation document, and text classification relates to efficiency and accurate rate. In this paper, a text classification model based on intelligence database is established firstly by the feature of military intelligence. Then the key technologies of text expression, automatic word segmentation, and feature extraction for the model were analyzed. Advanced algorithm of mutual information is proposed and simulated in military intelligence management. The result shows the method can extract normalization essential intelligence. This intelligence information can not only assist making battle decision project, but also be appended into intelligence database directly.

**Key words**: intelligence management; text classification; mutual information

## 0 引 言

近年来军事情报侦察的信息化程度不断提高,尤其注重对情报获取、处理等软硬件环境的建设,各情报侦察单元在长期的情报信息收集、处理、管理过程中,建立了大量的数据库用于保存原始数据和处理结果。对这些数据的有效处理及综合利用已经成为当前的一大难题,而在此基础上对情报数据库进行知识挖掘,则是当前情报研究中较为薄弱的环节。

笔者针对军事情报信息的特点,建立了一个基于情报数据库的文本分类模型,分析了模型中的文本表示、自动分词、特征提取等关键技术,并对其中的特征选取方法提出了改进措施,应用效果较好。

**收稿日期**:2006-11-07
**作者简介**:赵国际(1965-),男,北京人,博士,高工,研究方向为情报信息处理。

## 1 中文文本分类模型

### 1.1 文本的表示

目前,在中文信息处理中,文本的表示主要采用向量空间模型,简称 VSM。VSM 是 Salton 等[1]于 20 世纪 60 年代首先提出的,并在著名的 SMART 系统得到成功的应用。目前,该模型及相关技术,包括项的选择、加权策略,以及采用相关反馈进行优化查询等在文本分类、自动索引、信息检索等诸多领域得到广泛的应用,并取得了较好的效果。

下面定义向量空间模型的概念。

【定义 1】文档:泛指一般的文献或文献中的片断(段落、句子组或句子),一般指一篇文章。尽管文档可

以是多媒体对象,但以下讨论中只认为是文本对象,并对文本与文档不加区别。文本,用 $d$ 表示一篇文档(文本)。

【定义 2】项:文本的内容特征常常用它所含有的基本语言单位(字、词、词组或短语) 来表示,这些基本的语言单位被统称为文本的项,即文本可以用项集(Term List)表示为 $d(t_1, t_2, \cdots, t_n)$,其中 $t_k$ 是项,$1 \leqslant k \leqslant n$。

【定义 3】项的权重:对于含有 $n$ 个项的文本 $d(t_1, t_2, \cdots, t_n)$,常用一定的权重 $w_k$ 表示项 $t_k$ 在文本 $d$ 中的重要程度,即 $d = d(t_1, w_1; t_2, w_2; \cdots, t_n, w_n)$,简记为 $d = d(w_1, w_2, \cdots, w_n)$。

【定义 4】向量空间模型(VSM):忽略 $t_k$ 在文档 $d$ 中的先后顺序并要求 $t_k$ 互异,将文档 $d$ 简化以特征项的权重为分量的向量表示:$\boldsymbol{d} = \boldsymbol{d}(w_1, w_2, \cdots, w_n)$。即把 $t_1, t_2, \cdots, t_n$ 看成一个 $n$ 维的坐标系,而 $w_1, w_2, \cdots, w_n$ 为相应的坐标值,因而 $\boldsymbol{d}(w_1, w_2, \cdots, w_n)$ 被看成是 $n$ 维空间中的一个向量。称 $\boldsymbol{d}(w_1, w_2, \cdots, w_n)$ 为文本 $d$ 的向量表示。

【定义 5】相似度:对两个文本 $\boldsymbol{d}_1$ 和 $\boldsymbol{d}_2$ 之间的内容相关度(Degree of Relevance) 的度量被称为相似度 $\mathrm{Sim}(\boldsymbol{d}_1, \boldsymbol{d}_2)$。对于文档 $\boldsymbol{d}_1(w_{11}, w_{12}, \cdots, w_{1n})$ 和 $\boldsymbol{d}_2(w_{21}, w_{22}, \cdots, w_{2n})$,可以借助向量之间的某种距离来表示它们之间的相似度,常用向量之间的加权进行计算:

$$\mathrm{Sim}(\boldsymbol{d}_1, \boldsymbol{d}_2) = \sum_{k=1}^{n} w_{1k} * w_{2k} \tag{1}$$

或用夹角余弦值来表示:

$$\mathrm{Sim}(\boldsymbol{d}_1, \boldsymbol{d}_2) = \cos\theta = \frac{\sum_{k=1}^{n} w_{1k} * w_{2k}}{\sqrt{\sum_{k=1}^{n} w_{1k}^2 \sum_{k=1}^{n} w_{2k}^2}} \tag{2}$$

除了经典的 VSM 模型外,还有广义向量空间模型。传统的 VSM 中假设各词语间是相互独立的,无语义上的关系。广义向量空间模型,简称 GVSM,就利用文本而不是词来表示词之间的关系。其相似度的计算公式为:

$$\mathrm{Sim}(\boldsymbol{d}, \boldsymbol{q}) = \cos(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{d}, \boldsymbol{A}^{\mathrm{T}}\boldsymbol{q}) \tag{3}$$

其中 $\boldsymbol{A}$ 是 $m \times n$ 的文本矩阵,$m$ 是单词数,$n$ 是文本集中的文本数。

### 1.2　文本建模

对于内容这个难以表示的特征,首先要找到一种能够被计算机所处理的表示方法。向量空间模型(VSM)[2]是近年来应用较多且效果较好的方法之一。在进行文档表示前,首先进行文档的预处理,即对文档进行分词,得到一个词集合,对词进行词频统计,同时过滤停用词,剔除虚词,把出现频率低于一定范围的词排除掉。

根据一个文档集 $d$ 和一个项集合 $t$,可以将每个文档表示为在 $t$ 维空间 $R$ 中的一个向量 $v$。向量 $v$ 中第 $j$ 个数值就是相应文档中第 $j$ 个项的量度。如果文档不包括这个项,它就为 0,否则就不为 0。有许多定义这种非 0 权值入口的方法。例如:若第 $j$ 个项在文档中出现就简单地定义 $v_j = 1$ 或定义 $v_j$ 为项(在相应文档中) 出现的频数,或相对项频数,即项频数除以所有该项在相应文档中出现的次数。如表 1 所示,就是一个项频数矩阵。

表 1　项 - 文档词频矩阵

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|---|---|---|---|---|---|---|
| $t_1$ | 321 | 80 | 31 | 68 | 72 | 420 |
| $t_2$ | 354 | 91 | 71 | 56 | 82 | 298 |
| $t_3$ | 15 | 32 | 167 | 46 | 280 | 17 |
| $t_4$ | 22 | 142 | 78 | 200 | 52 | 15 |
| $t_5$ | 74 | 87 | 85 | 96 | 54 | 123 |

其中每列代表一个文档向量;每个入口 $(i, j)$ 记录项 $t_i$ 在文档 $d_j$ 中出现的次数。由于相似文档应具有相似的项频数,可以根据在频数矩阵中的内容,来判断一组文档间或文档与查询的要求间是否相似。

### 1.3　文本自动分词

基于 VSM 模型的文本分类中,关键的一步就是如何从文本中提取反映类别的有效特征。一般可以选择字、词或词组作为文本的特征,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。汉语文本不同于英文之处就在于,首先要将文本分词,才能进一步开展其它研究。

作为自然语言处理的前处理阶段,自动分词技术是重中之重,汉语自动分词是各种汉语信息处理包括语音处理、词频统计、主题词索引、文摘生成、情报检索、文本分类与聚类等工作的基础工程,也是制约中文信息处理飞跃的"瓶颈"之一。汉语的书面表达方式是以汉字为最小单位的,但是在自然语言理解中,词是最小的、能独立活动的、有意义的语言成分。把没有分割标志即没有词的边界的汉字串,转换到符合语言实际的词串即在书面汉语中建立词的边界,这就是汉语自动分词的任务。可以将现有的分词算法分为三大类:基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法[3]。

文中的分词系统采用的是基于字符串匹配的分词原理,这类分词系统通常是由分词预处理、切分串和分

词三部分组成，其原理如图 1 所示。

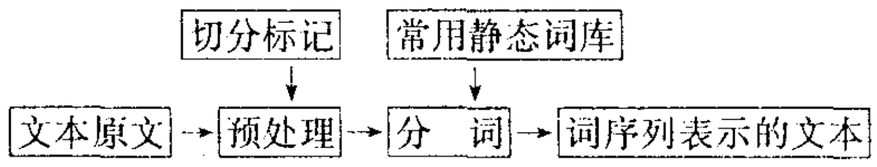


图 1 文本分词原理

对于输入的文本，利用切分标记（标点、数字、西文、其它非汉字符号）和隐式切分标记（出现在停用词表中的那些频率高、构词能力差的单词）将文本切分成汉字短串序列。

分词时，将经过分词预处理得到的短串用静态常用词词典对其进行匹配，得到文本的词序列文件。在此基础上进行特征提取工作。

### 1.4 文本特征选取

文本挖掘中一个重要的问题就是高维的特征空间，这些特征空间是由文本中的词或词组构成的，许多传统算法难以处理。高维特征集对于类别的区分能力来讲，未必全是重要和有益的，而且还会加剧计算的负担。在不影响特征分类准确度的情况下，减少文本描述空间的高维特征数量是很有必要的，这个过程称为特征选取（Feature Selection）。常用的特征提取方法有：文档频次（DF）、互信息（MI）、信息熵（IG）、$\chi^2$ 统计（CHI）等方法[4,5]。文中对互信息算法进行了改进，应用效果较好。

#### 1.4.1 互信息方法

互信息可以度量特征项和类别的共现关系，特征项对于类别的互信息越大，它们之间的共现概率也越大。假设文档集合 $C$ 分为 $K$ 类，记为 $C_1, C_2, \cdots, C_K$，特征项 $w$ 对于文档类别 $C_i$ 的互信息 $MI(w, C_i)$ 的计算公式如下：

$$MI(w, C_i) = \log \frac{P(w, C_i)}{P(w, C)} \tag{4}$$

其中 $P(w, C_i)$ 为特征项 $w$ 出现在类 $C_i$ 中的概率，$P(w, C)$ 为特征项 $w$ 在所有文档中的出现概率。

下面给出基于互信息的特征降维算法步骤：

(1) 如果特征项 $w$ 不是停用词，对于每一个文档类 $C_i, 1 \leqslant i \leqslant K$，计算 $w$ 对于 $C_i$ 的互信息 $MI(w, C_i)$

$$MI(w, C_i) = \log \frac{\mathrm{Freq}(w, C_i)}{\mathrm{Freq}(w, C)} \tag{5}$$

其中 $\mathrm{Freq}(w, C_i)$ 是 $w$ 在文档类别 $C_i$ 中出现的次数，$\mathrm{Freq}(w, C)$ 是 $w$ 在整个文档集 $C$ 中出现的次数。

(2) 特征项 $w$ 的对于文档集合 $C$ 的互信息 $MI(w, C) = MI(w, C_k)$，其中 $C_k$ 是特征项 $w$ 的互信息最大类别。

$$k = \underset{1 \leqslant i \leqslant K}{\mathrm{MAX}} \left( \frac{\mathrm{Freq}(w, C_i)}{\mathrm{Freq}(w, C)} \right) \tag{6}$$

(3) 给定阈值 $\alpha$，选择 $MI(w, C) > \alpha$ 的特征项组成特征空间。

#### 1.4.2 改进的互信息算法

特征选取算法的优劣直接影响到文本分类的效果，特征项选择依赖于频度、分散度和集中度等多项测试指标[2]。频度是最常用的特征选择测试指标，频度越高、分散度越大、集中度越强，则对文本分类越有用，即分辨度越强。

笔者提出了一种新的特征抽取 FS 算法，在互信息的特征抽取方法的基础上，给出分散度和集中度测试指标的修正，公式如下：

$$\mathrm{FS}(w, c_i) = \log_2 \left( \frac{P(w/c_i)}{P(w)} \right) \times \exp(n_i/n) = \log_2 \left( \frac{P(w, c_i)}{P(w) \times P(c_i)} \right) \times \exp(n_i/n) \tag{7}$$

其中 $n_i$ 为训练语料中特征项 $w$ 出现在类别 $c_i$ 中的文本数，$n$ 是训练语料中特征项 $w$ 出现的文本数，$P(w, c_i)$ 是训练语料中特征项 $w$ 出现在类别 $c_i$ 中的频率，$P(w)$ 是训练语料中特征项 $w$ 出现的频率。这样抽取出的特征能很好地体现频度、分散度和集中度测试指标，使其在这些指标中达到整体最优。

## 2 情报中心文本分类系统的设计

在未来侦察情报建设一体化框架下，情报基础数据库的数据量将是非常惊人的，并且积累速度也会非常快，对于这些数据，如果不能有效地加以利用，就会形成所谓的“数据坟墓”，成为无用的信息。而情报工作的主要任务是提供各种有价值的军事信息，为上级决策咨询服务。这就要求不但要提供及时、准确的情报信息，还要对历史的、实时的信息进行发掘，从中发现潜在的、隐含的知识，这正是数据挖掘要做的工作。基于此，提出了以情报服务为驱动的数据挖掘原型系统（MIDMS）[2]。

该系统从功能上可以分为三层：挖掘算法层、逻辑层、应用层，如图 2 所示。其最终目标是使得用户可以利用相关逻辑模型解决具体问题，直接获取所需的有价值的情报信息，而不必面对复杂的数据挖掘算法。

| | | | | |
|---|---|---|---|---|
| 应用层 | 军事情报信息系统 | 军用信息网、科研训练网、因特网 | 遥感影像、卫星资料管理系统 | … |
| 逻辑层 | 文本情报数据挖掘 | WEB 数据挖掘 | 空间数据挖掘 | 系统逻辑 |
| | 情报分类、主题提取、趋势分析… | 网站结构优化、网页推荐、网站监控… | 目标识别、毁伤评估、景象匹配… | 系统模型 |
| 挖掘算法层 | 关联规则、序列模式、时间序列、分类、聚类、异常检测… | | | |

图 2 军事情报数据挖掘系统层次结构

依据上述的功能划分，所设计的原型系统体系结构如图 3 所示，该系统从左到右依次为三个层面：

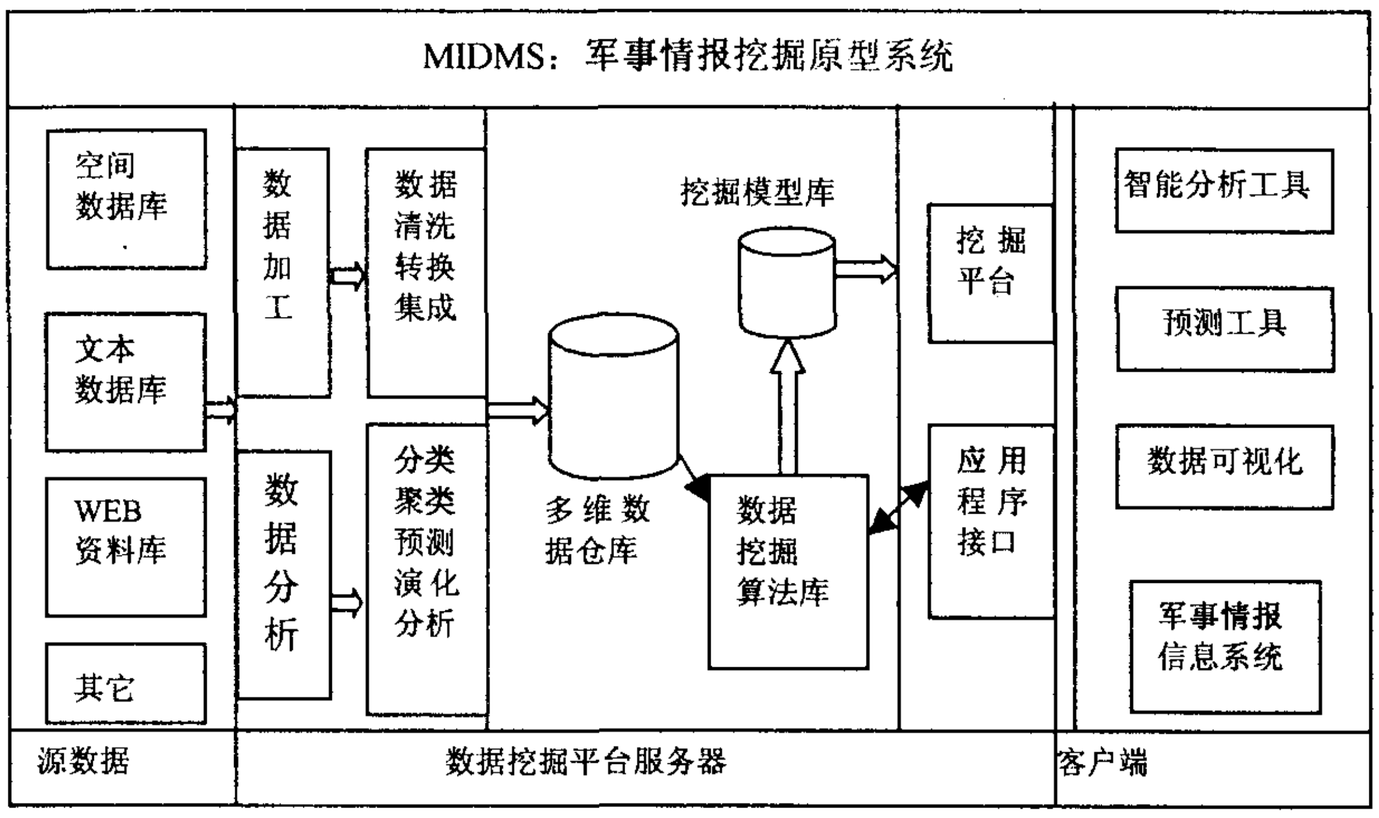


图 3　军事情报数据挖掘系统原型

第一层主要是信息系统,包括空间数据库、文本数据库及 WEB 资料库,具体一点,主要是指情报信息系统中的目标库、预警信息库、遥感影像库、侦察实力库、动向情报库、声像资料库等,该层向数据挖掘平台提供原始数据。

第二层是挖掘平台的应用服务器层,该层首先将第一层的原始数据经过抽取、转换、清洗,装载到多维数据仓库中,然后以此为基础,通过数据挖掘算法,定时分析经过整合后的数据,将挖掘的结果存放在挖掘模型库中,应用服务器向客户端提供各模型算法的 API 函数接口,便于二次开发。

第三层是挖掘平台的客户端软件,这些软件基于全军一体平台开发,直接面向用户,用于对其数据的智能分析、预测及实现数据的可视化。

## 3　结束语

着眼于情报数据库中的文本信息分类,针对现在存在的问题,对文本分类的关键技术进行了研究,并提出了改进的文本特征提取方法。在以情报服务为驱动的数据挖掘原型系统的基础上,设计了原型系统体系结构,效果较好。

**参考文献:**

[1] Salton G, Lest M E. Computer Evaluation of Indexing and Text Processing[J]. Association for Computing Machinery, 1968,15(1):8－36.

[2] 闫万体,赵宗涛,陈晓峰.数据挖掘及其在军事情报系统中的应用研究[J].自动化指挥与计算机,2005(3):32－35.

[3] 湛　燕,陈　昊,袁　方,等.基于中文文本分类的分词方法研究[J].计算机工程与应用,2003,39(23):97－91.

[4] 胡佳妮,徐蔚然,郭　军,等.中文文本分类中的特征选择算法研究[J].光通信研究,2005,3:44－46.

[5] 闫万体.文本分类技术及在二炮情报分析中的应用[D].西安:第二炮兵工程学院,2004.

(上接第 132 页)

现出降低—快速降低—较缓慢降低的过程,这种变化过程在维度较小的情况下尤其明显。

表 1　性能评估表

| 实视图个数 | 存储空间(kB) | 查询所用维度个数 | 查询响应时间(ms) |
|---|---|---|---|
| 0 | 28320 | 2 | 94632 |
| | | 3 | 112775 |
| 5 | 83978 | 2 | 88110 |
| | | 3 | 103486 |
| 10 | 150211 | 2 | 58274 |
| | | 3 | 81230 |
| 15 | 214909 | 2 | 27104 |
| | | 3 | 73972 |
| 20 | 312475 | 2 | 25166 |
| | | 3 | 67330 |

当实视图数量较少时,ROLAP 查询响应时间无明显改善。同时,当实视图数量达到一定数目后,再增加实视图数目对提高查询效率的作用也越来越小。由此可见添加了一定数量的实视图并采用改进的 VSP 算法提供 OLAP 服务,虽然存储空间随着实视图个数的增加有所上升,但是平均相应时间却有明显改善。因此只要设定恰当的实视图的个数,在存储空间和查询效率之间取一个恰当的平衡点,完全可以在系统资源允许的范围内提供高效的 OLAP 服务。

**参考文献:**

[1] 王能斌,董逸生. 数据库设计与实现[M].武汉:华中理工大学出版社,1991.

[2] 杭晓骏.一个通用 OLAP 建模工具的设计与实现[D]. 南京:东南大学计算机系,2005.

[3] 谭红星, 周龙骧. 多维数据实视图的动态选择[J]. 软件学报, 2002, 13(6):1090－1096.

[4] Nadeau T P. Achieving Scalability in OLAP Materialized View Selection [D]. Michigan: America University of Michigan, 2002.

[5] Gunderloy M, Sneath T. SQL Server 开发指南: OLAP(联机分析处理)[M].北京:电子工业出版社,2001.