

# 文本聚类分析在故障诊断中的应用

张霞,王建东,庄毅,邹玉娟

(南京航空航天大学 信息科学与技术学院,江苏 南京 210006)

**摘要** 针对自然语言理解困难以及小样本文本分类的技术难点,提出了基于聚类分析的降落伞故障诊断新方法。利用最大树法来实现对小样本案例的聚类与提取,避免了制定推理规则的复杂性。通过“降落伞建模验模与设计优化系统”中的实例分析,说明了所提出的基于文本聚类分析的故障诊断方法有效可行。

**关键词** 故障诊断;聚类分析;最大树法

中图分类号:TP182

文献标识码:A

文章编号:1673-629X(2007)02-0008-04

## Application of Clustering Analysis of Text in Fault Diagnosis

ZHANG Xia, WANG Jian-dong, ZHUANG Yi, ZOU Yu-juan

(College of Computer Sci. and Techn., Nanjing Univ. of Aeronautics and Astronautics, Nanjing 210006, China)

**Abstract** Proposes a new method of parachute fault diagnosing based cluster analysis, which aims at the difficulty of language understanding and categorization for small sample. Using maximal tree method to cluster and extract the small cases, it avoids the complexity of establishing reasoning rules. Experimental results show that the approach is effective and feasible by using it in Parachute Modeling and Validating and Design Optimization System.

**Key words** fault diagnosis; clustering analysis; maximal tree method

## 0 引言

现代化生产中,很多科学领域并不是精确性的科学或者还没有形成一套精确的理论依据,例如降落伞设计理论就是一个典型的例子,其投放实验出现的故障经常要依靠经验来找出原因,而且也很难对这种原因或解决方案做出令人信服的解释,只是凭借专家的经验知识来确定,因此迫切需要人为地“保存”和利用好这些宝贵的经验。

文本聚类分析,顾名思义,即对文本用聚类法分析,从而发现其有用的模式和知识,该技术已经成为文本挖掘中一个日益流行而且重要的研究领域。而专家的经验知识直接表示形式为文本,故该技术能很好地为非精确性科学领域服务。

聚类分析的方法很多,文中着重讨论模糊相似矩阵用最大树来完成聚类过程,提出并详细阐述了在“降落伞建模验模与设计优化系统”中,针对小样本文本聚类分析的故障诊断的应用与实现。

## 1 基于案例的故障诊断

故障诊断是随着生产过程的复杂化而产生的一种技术,而在一些设计复杂的产品领域中,并不可以精确测到每个过程的每个参数的精确值,所以当出现故障时找其原因是一件很难的事情。例如降落伞空投试验中经常会出现一些故障,这时需要依靠该领域的专家凭借丰富的知识和经验进行故障的诊断与排除工作。

运用专家系统中基于案例推理的思想,在此类特殊领域中的故障诊断可以把以前使用过的与当前问题类似的案例联系起来,通过访问过去相似问题的解决方法而获得当前问题求解方法;当接受一个求解新问题的要求后,利用相似度知识从案例库中找出与当前问题最相关的案例。考虑到从专业领域专家那得到的一般都是自然语言形式的知识,特别是像降落伞等领域的知识很难用形式化的规则描述,故尝试寻求一种可以对此类“一手”知识进行处理和分析的方法,而文本聚类分析技术可以很好地运用到此类“原始”知识上,用以发现一些有用的信息并应用到生产实际中。

## 2 文本聚类分析技术

### 2.1 聚类分析概述

聚类分析(Clustering Analysis)是文本挖掘中的重

收稿日期:2006-05-09

基金项目:航空十五预研支撑项目(41801150201)

作者简介:张霞(1981-),女,江苏人,硕士研究生,研究方向为知识工程与机器学习;王建东,教授,博士生导师,主要从事数据挖掘、知识工程与机器学习方面的研究。

要技术之一,针对没有预先定义主题类别的文本来说,采用聚类分析是个良策。具体来讲,它是指将文档集合分成若干个簇,要求同一簇内文档内容的相似度尽可能的大,而不同簇间的相似度尽可能的小,从而发现整个文本集合的整体分布特点<sup>[1]</sup>。

2.2 文本聚类分析一般流程

文本聚类分析的一般流程为:文本原文→预处理→分词→特征项表示→模式或知识的提取→模式或知识的运用。

- (1) 选取待处理和分析的文本;
- (2) 对得到的文本进行预处理:利用切分标记(标点、数字等)和隐式切分标记(出现在停用词表中的那些频率高、构词能力差的词,如的、了)将文本切分成短串序列;
- (3) 对预处理后的文本进行分词处理;
- (4) 把文本切分成特征词条,建立挖掘对象的特征表示,一般采用文本特征向量,若维数过大还需进行降维处理;
- (5) 利用聚类分析相关技术来提取潜在的模式或知识;
- (6) 运用提取的模式或知识。

2.3 模糊聚类分析方法

模糊聚类分析方法是 将模糊概念的方法和理论引入聚类分析,主要需要进行系数标定、聚类分析等步骤。系数标定是指计算出两个样本  $i$  与  $j$  之间的相似程度,当这个值越接近于 1 时,表明这两个样本越接近。聚类分析是利用最大树法等对相似矩阵进行自动分类,并可从结果中得到一些信息。

3 基于最大树聚类分析的降落伞故障诊断

3.1 设计思想

鉴于前文所述降落伞领域非精确性的特点,针对从专家处得到的都是文字性的故障描述,故笔者从基于案例的角度出发,利用文本聚类分析技术从这些宝贵的“描述”中得到一些有用的信息。

要很好地管理专家的经验知识,很重要的一步就是对此类文本信息的分类。而由于样本数小等原因不能实现对文本的训练,故需寻求一种适用的无指导学习方法,分类伴随着模糊性,将模糊数学中的有关概念与方法引进聚类分析,通过建立模糊相似关系(模糊相似矩阵)之后直接进行分类。相应地,吴望名提出了直观意义很明确的“最大树”方法<sup>[2]</sup>,而对于仅具有自反性和对称性的模糊相似关系的矩阵可用最大树法直接分类<sup>[3]</sup>。基于这些考虑,笔者灵活地把此方法用在了文本聚类分析中。

3.2 最大树聚类法以及在聚类分析中的应用

(1) 最大树聚类法。

文中探讨的最大树聚类法主要是利用模糊相似矩阵来完成对事物聚类分析,那如何构造一棵最大树(“赋权树”)并实现聚类是关键。模拟图论中“最小生成树”的概念:它是一个连通图的极小连通子图,它包含原图中的所有顶点,而且有尽可能少的边<sup>[4]</sup>。这意味着对于最小生成树来说,若砍去它的一条边,就会使该生成树变成非连通图,从而形成不同的簇。文中采用类似的方法来构造最大树并实现聚类。

根据生成树的定义,为得到最大生成树,人们设计了很多算法,最著名的有 prim 算法和 kruskal 算法,两者都采用了一种逐步求解(Crady)的策略:设有一个连通网络  $N = \{V, E\}$ , 顶点集合  $V$  中有  $n$  个顶点,最初先构造一个包括全部  $n$  个顶点和 0 条边的森林  $F = \{T_0, T_1, \dots, T_{n-1}\}$ ,以后每一步向  $F$  中加入一条边,它应当是一端在  $F$  的某一棵树  $T_i$  上,而另一端不在  $T_i$  上的所有边中具有最大权值的边。边的加入使得  $F$  中某两棵树合并为一棵,树的棵数减一。经过  $n - 1$  步,最终得到一棵有  $n - 1$  条边的各边权重值总和达到最大的最大生成树<sup>[5]</sup>。

文中采用了 Prim 算法来实现,其基本思想是:从连通网络  $N = \{V, E\}$  中的某一顶点出发,选择与它关联的具有最大权值的边,将其顶点加入到生成树的顶点集合中。同时,文中的最大树法是在得到的最大生成树基础上,选定某阈值进行剪枝,从而分裂成若干棵子树,而相应地一棵子树对应着一类。下面给出 Prim 算法 vc 的实现:

```
vector <edge> tree //存放最大生成树
vector <float> lowcost //存放生成树顶点集合内顶点到生成树外各顶点的边上的当前最大权值

vector <int> nearvex ;
//记录生成树顶点集合外各顶点距离集合内哪个顶点最相似

tree.clear();
lowcost.clear();
nearvex.clear();
lowcost.push_back(0);
nearvex.push_back(-1);
for (int p=1 ;p<m ;p++ )/m 为顶点总数
{
    lowcost.push_back( weightal[ 0 ][ p ] );
    // weightal[ i ][ j ] 为顶点 i 与顶点 j 的相关度
    nearvex.push_back(0);
}
edge e ;
// struct edge//最大树边的存储结构
```

```

// {
// int beg ;
// int end ;
// float weight ;
// int clas ;
// };
for ( p = 1 ; p < m ; p ++ )
{ //求生成树外顶点到生成树内顶点具有最大权值的边
    float min = 0.0001 ;
    int v = 0 ;
    for ( int q = 0 ; q < m ; q ++ )
    { //确定当前具最大权值的边及顶点位置
        if ( nearvex[ q ] != - 1 && lowcos[ q ] > min )
        {
            v = q ;
            min = lowcos[ q ] ;
        }
    }
    if ( v / v = 0 表示再也找不到要求的顶点了
    {
        e. end = nearvex[ v ] ;
        e. beg = v ;+ e. weight = lowcos[ v ] ;
        e. clas = 0 ;
        tree. push_ back( e ) ;
        nearvex[ v ] = - 1 //加入生成树顶点集合
    }
    for ( q = 1 ; q < m ; q ++ )
    {
        if ( nearvex[ q ] != - 1 && weightal[ v ][ q ] > lowcos[ q ] )
        { lowcos[ q ] = weightal[ v ][ q ] ;
          nearvex[ q ] = v //修改 }
        }
    }
}
//顶点表示案例 ID 号 ,边表示案例间的相似度

```

## (2) 基于最大树聚类分析法的降落伞故障诊断。

根据前述文本聚类的一般流程 ,相应地 ,文中基于最大树聚类分析法的降落伞故障诊断的流程为 :

第一步 :预处理。对获得的降落伞故障描述去除标点符号处理(故障描述可有重复)。

第二步 :故障案例的表示。利用常用静态词库对故障案例(文本形式)向量表示 ,一条记录对应成一个若干维的向量 ,即得到一个记录/词条矩阵。

第三步 :生成模糊相似矩阵。依据记录/词条矩阵得到每条记录之间的两两相似度关系 ,得到故障案例之间的模糊相似矩阵。

第四步 :自动分类。利用最大树法在模糊相似矩阵基础上实现自动分类(见上文所述) ,得到若干类的故障。

第五步 :利用分类信息。对于新请求的故障 ,找出

最相似的案例记录 ,并返回与该案例同类的所有记录给用户。

### 3.3 具体实现过程

基于前文所述 ,提出基于最大树聚类分析法的故障诊断 ,对现有知识库聚  $n$  个故障类并随案例库动态的改变而变化。具体实现如下 :

Step 1 :建常用静态词库 ,只要可以构成词语并含有含义都加入词库 ,比如“碰撞”、“有碰撞”以及“无碰撞”都作为有意义的词入词库。

Step 2 :取出以前的故障案例来分析 ,每条记录都用一个  $N$  维的布尔向量来表示(  $N$  指词库中关键词的数目) ,并临时保存此动态的向量矩阵  $M * N$  (  $M$  为案例条数)。

Step 3 :对 Step 2 得到的布尔型矩阵求故障案例之间的两两相似度 ,即通过求不同案例(针对相同关键词)的同或关系得到不同案例之间两两相似度得到相似矩阵  $M * M$ 。

Step 4 :用简单相关系数法(又称 PEARSON 相关系数法)对 Step 3 得到的矩阵标准化 ,相关度比较容易分析 ,如标准化后 item1 和 item1 的相关度就是 1。所用公式如下 :

$$r_{ij} = \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^m (x_{ik} - \bar{x}_i)^2} \cdot \sqrt{\sum_{k=1}^m (x_{jk} - \bar{x}_j)^2}}$$

$$\text{式中 } \bar{x}_i = \frac{1}{m} \sum_{k=1}^m x_{ik}, \bar{x}_j = \frac{1}{m} \sum_{k=1}^m x_{jk}。$$

这里  $r_{ij}$  表示两个案例  $x_i$  与  $x_j$  之间相似程度的变量 ,当  $r_{ij}$  接近于 1 ,表明这两个样本越接近。

Step 5 :用 prim 算法求最大生成树 ,节点为案例标号 ,路径标上计算出来的标准权重(具体的实现过程见上文)。

prim 算法的基本思想为 :

(1) 先在  $M$  中任取一个案例节点  $v_0$  (之后都称节点) ,并取入  $T$  中 ;

(2) 令  $S = V(G) \setminus V(T)$  ,其中  $V(G)$  ,  $V(T)$  分别为  $G$  与  $T$  的节点集 ;

(3) 在所有连接  $V(T)$  的节点与  $S$  的节点的边中 ,选出权数最大的边(  $u_0, v_0$  ) ;

(4) 将边(  $u_0, v_0$  )取入  $T$  中。重复(2)至(4)步骤 ,直至  $G$  中的节点全都取入  $T$  中为止。

Step 6 :选定阈值对于 Step5 求得的树开始分裂(聚类分析) ,若选定阈值为  $\lambda$  ,则断开权重  $w < \lambda$  的路径得到 kind 棵子树 ,而每棵子树就是聚成的一个故障类 ,并记录每个案例所属的类。

Step 7 :对于一个用户故障诊断要求 ,按照前面所

述方法对此请求用向量表示 ,再与数据库中案例比较求相似度再标准化 ,得到与数据库中第  $i$  条案例最相似并取得它的类标号(若出现若干条案例相同则任取一个  $i$ )。

Step 8 取得 Step7 得到的类标号 ,同时按相似度从大到小反馈给用户此类标号的案例情况(包括故障原因、解决方案等)。

3.4 实验数据

目前从专家那收集到的故障案例中共有 31 条记录 ,若记录扩充后再自动重新聚类 ,形成的最大树示意图如图 1 所示(其中节点 0~30 表示 31 个案例 ,树枝表示两端点间的相似度 ,取  $\lambda = 0.900$  ,得到 kind = 8)。

本实验中形成的最大树不仅可得到合适的分类 ,而且还可以得到更多的潜在信息 ,如 :某两个节点所在类(标号)差别越大说明此两个案例的差别越大 ;不同的  $\lambda$  得到不同要求的分类等等。若不需要把案例分的很清晰的话 ,可通过降低  $\lambda$  值来实现 ,如  $\lambda = 0.5$  可分为 :

《0,1,2,7,8,9,10,11,26,29》《12,13,14,15,16,30,3,4,5,6》《17,22,27,21,18,19,28,23,24,25》《20》四个类 ,故需要寻求一个合适的  $\lambda$  值以取得用户需要的精度的分类。

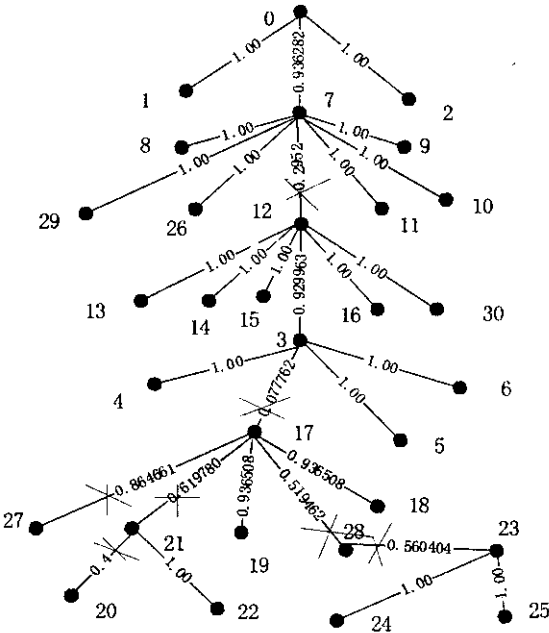


图 1 最大生成树

根据最大树得到的降落伞故障案例类别标注如表 1 所示(注 :同一故障会有几种不同的原因或解决方案 ,库表中有一些故障是相同的 ,所以出现很多项相关度为 1 的情况) ,从表中发现结果符合人们直观上的分类。

4 总结与展望

由于自然语言理解技术本身存在问题 ,而从领域专家的理解所描述的知识然后制定推理规则 ,那将又多了道槛。文中从文本聚类分析角度出发 ,提出对专家的文字描述进行整理和聚类分析 ,为后人查询专家的经验提供帮助 ,实验表明 ,这是条有效路径。而领域知识库达到一定规模的话 ,可以有更好的聚类分析结果 ,比如 (1)根据聚类分析可从特征文本中抽取一些作为索引项并命名 ,用户可以直观地知道案例库有哪些种类的知识 (2)产生比较好的聚类中心 ,可为用户提供更快速的查询。

表 1 分好类的故障描述

ID	故障描述	所属类
0	单个伞衣有灼伤 ,伞衣间无碰撞	1
1	单个伞衣有灼伤 ,伞衣间无碰撞	1
2	单个伞衣有灼伤 ,伞衣间无碰撞	1
3	多个伞衣有灼伤 ,伞衣间无碰撞	2
4	多个伞衣有灼伤 ,伞衣间无碰撞	2
5	多个伞衣有灼伤 ,伞衣间无碰撞	2
6	多个伞衣有灼伤 ,伞衣间无碰撞	2
7	单个伞衣有灼伤破损 ,伞衣间无碰撞	1
8	单个伞衣有灼伤破损 ,伞衣间无碰撞	1
9	单个伞衣有灼伤破损 ,伞衣间无碰撞	1
10	单个伞衣有灼伤破损 ,伞衣间无碰撞	1
11	单个伞衣有灼伤破损 ,伞衣间无碰撞	1
12	多个伞衣有灼伤破损 ,伞衣间有碰撞	2
13	多个伞衣有灼伤破损 ,伞衣间有碰撞	2
14	多个伞衣有灼伤破损 ,伞衣间有碰撞	2
15	多个伞衣有灼伤破损 ,伞衣间有碰撞	2
16	多个伞衣有灼伤破损 ,伞衣间有碰撞	2
17	伞衣塌陷	3
18	伞衣冲破	3
19	伞衣有油污	3
20	伞衣套可卸橡皮绳套圈损伤	8
21	伞衣错位引起操纵纵带错位	5
22	伞衣错位引起操纵纵带错位	5
23	稳降时间达不到要求	7
24	稳降时间达不到要求	7
25	稳降时间达不到要求	7
26	单个伞衣有灼伤破损 ,伞衣间无碰撞	1
27	伞衣漂离问题	4
28	开伞同步性不好	6
29	单个伞衣有灼伤破损 ,伞衣间无碰撞	1
30	多个伞衣有灼伤破损 ,伞衣间有碰撞	2

目前大多数信息均表现为文本方式 ,文本挖掘技术随着信息时代的发展将日益重要 ,若能真正做到从文本数据中提取信息 ,从信息中及时地发现知识 ,将可以大大地为人类的思想决策和战略发展服务。

参考文献 :

[1] Fu Weipeng. Text Document Clustering and the Space of Concept on Text Document Automatically Generated[D]. Graduate school of University of Science and Technology of China , 2001.



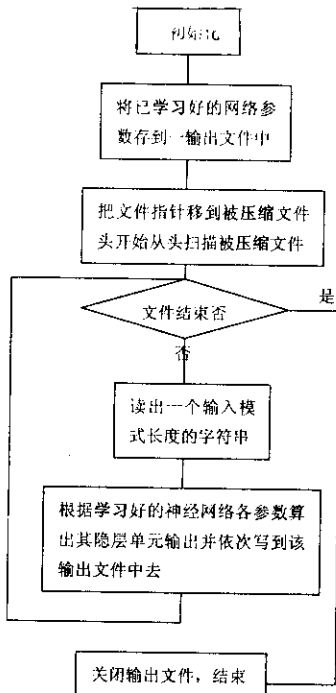


图 3 压缩流程图

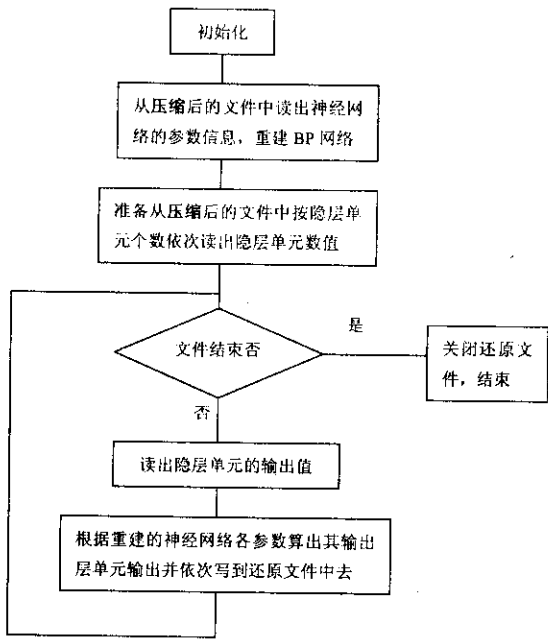


图 4 解压流程图

② WHILE NOT eof( targetfile )

```
BEGIN
Thide= readhideunitva( );
Tout= computetoooutput( Thide );
Writetoretfile( Tout ,retfile );
```

END ;

从被压缩文件( targetfile )中还原( readhideunitval )出隐层单元值 Thide ,再根据重建的 BP 网络得到与输入模式一致的输出 Tout( computetoooutput ) ,再将其写入到解压文件( retfile )中去( Writetoretfile )。此处 , Thide ,Tout 均为值集。

5 关于再压缩

从理论上说 ,按新思路的方法可以重复多次压缩信息 ,以达到理想的压缩比 ,大大减少信息传播所需的时间和空间。但因多次压缩需经多次解压才能还原 ,且每次压缩都可能产生信息缺损 ,故应注意压缩次数不要超过一定限度 ,以避免压缩时间、解压时间和信息缺损增加到不可忍受的地步。

6 结束语

用神经网络的方法实现数据压缩的新思路 ,这本身就是一个突破。它必将带动一个新领域研究的发展。这种方法也促使人们更多地考虑到对已有压缩成果的再压缩 ,对压缩极限有了新的认识 ,开创了压缩理论的新天地。数据压缩技术是现代信息社会所急需的热门技术之一 ,仍然还有很多的方面值得人们去研究。现在 ,不仅是上述的方法 ,很多新理论、新技术也层出不穷。小波变换、分形压缩等等已经成为数据压缩的新热点。上面提到的与神经网络结合的数据压缩技术也将占有一席之地 ,并将以此为基础 ,不断促进信息处理的智能化、自动化。

参考文献 :

[ 1 ] 吴乐南. 数据压缩的原理与应用[ M ]. 北京 :电子工业出版社 ,1996.

[ 2 ] 焦李成. 神经网络系统理论[ M ]. 西安 :西安电子科技大学出版社 ,1996.

[ 3 ] 王彦春 ,余钦范 ,段云卿. 改进的快速 BP 算法[ J ]. 物探与化探 ,1999 ,23( 2 ) :133 - 137.

[ 4 ] Werbos P J. Back Propagation Through Time :What It Does and How to Do It[ J ]. Proc. of the IEEE ,1990 ,78( 10 ) :1550 - 1560.

[ 5 ] 侯 阳. 数据压缩技术与 C 语言实例[ M ]. 北京 :学苑出版社 ,1994.

(上接第 11 页)

[ 2 ] 吴望名. 弗晰图与弗晰树[ J ]. 数学的实践与认识 ,1980 ( 4 ) :13 - 16.

[ 3 ] 王耀南 ,李树涛. 计算机图像处理与识别技术[ M ]. 北京 :高等教育出版社 ,2001.

[ 4 ] 杨炳儒. 图论概要[ M ]. 天津 :天津科学出版社 ,1990.

[ 5 ] 严蔚敏 ,吴伟民. 数据结构[ M ]. 北京 :清华大学出版社 ,1992.