第34卷 第19期 计 算 机 工 程 2008年10月
Vol.34 No.19 Computer Engineering October 2008
·软件技术与数据库· 文章编号：1000—3428(2008)19—0064—03 文献标识码：A 中图分类号：TP393.09

# 基于潜在语义差异的医学网页聚类

米晓芳，秦 洋，王立宏，宋宜斌
(烟台大学计算机学院，烟台 264005)

**摘 要：**采用潜在语义索引的全局模型和局部模型表示医学网页时，模糊聚类结果的类间包含度很大。该文提出一种新的潜在语义差异模型，将医学网页中的文本抽取出来并分别采用全局模型、局部模型和差异模型进行表示，利用FCM算法进行聚类并计算类间包含度。实验发现，对给定的5类医学网页进行聚类时，采用差异模型时的类间包含度平均约为全局模型的85%、局部模型的80%。

**关键词：**潜在语义索引；差异模型；文本挖掘；FCM聚类；包含度

# Medical Webpage Clustering Based on Latent Semantic Difference

MI Xiao-fang, QIN Yang, WANG Li-hong, SONG Yi-bin
(College of Computer, Yantai University, Yantai 264005)

**【Abstract】** Fuzzy clustering, two categories of medical Web pages represented by global LSI or local LSI generate two fuzzy sets with a large inclusion degree. A new latent semantic difference model is proposed. The text in medical Webpage is extracted and represented by global LSI, local LSI and difference LSI respectively. FCM algorithm is employed to cluster the feature vectors and inclusion degree between two result fuzzy sets is calculated. Experiments on five given categories of medical Webpages show that, on the average, difference LSI reduces the inclusion degree by a factor of 85% and 80% respectively when compared with global LSI and local LSI.

**【Key words】** latent semantic index; difference model; text mining; FCM clustering; inclusion degree

## 1 概述

随着Internet的发展，互联网出现了海量的、异构的Web信息资源，其中，Web文本占了主导地位。如何从这些海量的Web文本信息中获得有价值的信息，成为信息处理领域的一个关键问题。人们将数据挖掘技术应用到Web的知识发现中，形成了Web挖掘技术。

循证医学网页记录着丰富的循证医学实验信息，涉及疾病防治、保健、药物等各方面，典型的网上数据库有：Cochrane, Evidence-Based Medicine, Medline等，用户可以免费下载各种报告文章的摘要信息。文献[1]以Medline中的摘要信息为研究对象，抽取出病人群体、对比治疗、疗效评价等信息。本文研究这些医学网页的聚类问题，主要考查样本网页聚类中涉及的文本表示与降维问题。

潜在语义索引挖掘文本与特征之间潜在的高阶语义结构，将分解文本特征矩阵，实现特征空间的降维，文本和特征被转换到低阶语义空间上进行描述。目前主要的潜在语义模型有全局模型及其改进模型、局部模型及其改进模型[2]。本文在现有的全局模型和局部模型基础上提出一种潜在语义差异模型。通过对Medline数据库下载的医学网页数据进行实验，结果表明：利用该模型表示文本向量能有效地改进类间包含度。

## 2 潜在语义索引差异模型

潜在语义索引(Latent Semantic Indexing, LSI)[3-4]的基本思想为：文本中的词与词之间存在某种联系，即存在某种潜在的语义结构，可以采用统计的方法来寻找该语义结构，并且用语义结构来表示词和文本。这样可以消除词之间的相关性，化简文本向量。

潜在语义分析是基于矩阵的奇异值分解(Single Value Decomposition, SVD)技术的。给定一组文档，假定包含$n$篇文档，其中有$m$个不同的词项，采用词项作为特征。该文档集可以表示为词项×文档矩阵：$\boldsymbol{X}=[x_{ij}]_{m\times n}\in R^{m\times n}$。其中，矩阵元素$x_{ij}$表示词项$t_i$在文档$d_j$中的权值。矩阵$\boldsymbol{X}$的奇异值分解可以表示为

$$\boldsymbol{X}=\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} \tag{1}$$

其中，$\boldsymbol{U}$和$\boldsymbol{V}$分别为左奇异矩阵和右奇异矩阵；$\boldsymbol{U}$是一个$m\times m$的列标准化正交矩阵；$\boldsymbol{D}$是一个$m\times n$的对角矩阵，其对角线上的元素是按降序排列的非负奇异值；$\boldsymbol{V}$是一个$n\times n$的正交矩阵。

选择适当$k$值，将$\boldsymbol{D}$中删除相应的行和列得到$\boldsymbol{D}_k$，删除$\boldsymbol{U}$,$\boldsymbol{V}$相应的行和列分别得到$\boldsymbol{U}_k$,$\boldsymbol{V}_k$，得到新的矩阵$\boldsymbol{X}_k=\boldsymbol{U}_k\boldsymbol{D}_k\boldsymbol{V}_k^{\mathrm{T}}$。这个新矩阵是对原矩阵的逼近，可以用它去近似原始矩阵。LSI空间是由$\boldsymbol{U}_k$的前$k$维列向量张成的空间，因此，矩阵$\boldsymbol{X}$中的文档可以投影到$k$维LSI空间，得到低维表示。LSI不但可以对文本表示进行有效的降维，还可以捕捉到文本的潜在语义信息[2]。

目前主要的潜在语义模型有全局模型、局部模型两种[2]。假设有两个类，它们的特征向量分别用$\boldsymbol{A}$和$\boldsymbol{B}$表示(注：这里的向量是特征词项的集合，以下相同)。对于全局的潜在语义模型来说，它的特征向量是取$\boldsymbol{A}$和$\boldsymbol{B}$的并集，即$\boldsymbol{A}\cup\boldsymbol{B}$组成全局的特征向量，然后用得到的特征向量对两类的文本集合求词项×文档矩阵。对于局部的潜在语义模型来说，它的特征向

**基金项目：**国家自然科学基金资助项目(60473115)；山东省自然科学基金资助项目(Y2006G22)

**作者简介：**米晓芳(1982－)，女，硕士研究生，主研方向：Web信息挖掘；秦 洋，硕士研究生；王立宏、宋宜斌，教授

**收稿日期：**2007-11-23 **E-mail：**xiaofangmi@yahoo.com.cn

量是取$A$或者$B$的特征向量来表示两类的文本集合的矩阵。全局模型将两类的特征混合起来，综合了两类的特征，但不利于表示类之间的差异；而局部模型用一个类的特征表示两个类，有些重要的特征词项就会无法表示出来，从而影响到两类之间的区分。

本文提出一种新的潜在语义索引差异模型，采用两类的特征向量 $A$ 和 $B$ 的对称差来表示新的特征向量 $C$，即

$$C = A \oplus B = (A - B) \cup (B - A) \tag{2}$$

差异模型保留了两类之间的不同点，既保留了 $A$ 不同于 $B$ 的特征，同时也保留了 $B$ 不同于 $A$ 的特征。可以期望不同类的样本采用该模型表示时能有较大差异，从而使类之间的边界更清晰。该模型处理多个类别时可以采用“1 对其余”的策略。后续实验在对网页进行聚类时发现：采用差异模型时类之间的包含度比采用全局模型和局部模型都有下降，从而证实了差异模型是有效的。

## 3 医学网页聚类

### 3.1 医学网页预处理

Web 挖掘是在文本挖掘技术上发展起来的，由于网页本身的特征使得 Web 挖掘比文本挖掘更加难处理，因此本文首先将网页转化为文本文档，再利用文本挖掘技术对其进行处理。实验数据是从 Medline 医学数据库中按类别下载的英文医学网页，由于是从同一个数据库上下载的，因此网页的结构基本上是相同的。通过分析发现，网页中正文是在标签<D>和</D>之间的，从<D>开始“文章标题”处于第一个<B>和</B>标签之间，“文章作者”处于第 2 个<B>和</B>标签之间，紧接着“摘要正文”处于<BR><BR>和<BR><BR>标签之间。可以通过 C 语言编程读出网页中的标题、作者和摘要，并将其保存为与网页同名的文本文档。

在对文档进行特征提取前，需要对文本信息进行预处理。主要完成以下几方面工作：

(1)建立英文停止词表(stop word)，这种词几乎不携带任何信息，如“the”,“a”等连词和虚词对表示网页特征没有意义。

(2)统计每一篇文档的词项和词频，汇总词项得出每一类的特征向量($term_1$,$term_2$,…, $term_n$)。以类别为单位，计算每个特征词项$term_j$出现的文档个数，即文档频率。

(3)删除一些低频词，比如文档频率 < 5 的单词，这样就得到了每个类的频繁词特征向量，这是真正使用的特征向量。

### 3.2 FCM 算法和包含度

模糊$c$-均值(FCM)算法是模式识别的一个重要工具，FCM聚类算法适用于每类样本数相差不大且每类样本点距离相差不大的团状数据。FCM算法需指定数据集的类别数目$c$，本文只讨论两类之间的聚类问题，故取$c$=2。模糊聚类问题可表示为以下数学问题[5]：

$$\min\ J_m(\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} \quad \text{s.t.} \sum_{i=1}^{c} u_{ik} = 1, \sum_{k=1}^{n} u_{ik} > 0, u_{ik} \geq 0$$

其中，$u_{ik}$是第$k$个样本点对第$i$个类的隶属度，$k$=1,2,…,$n$；$\boldsymbol{U}$=$[u_{ik}]_{c\times n}$是隶属度矩阵；$m$通常取 2。

FCM 算法的实现步骤如下：

(1)初始化。随机给出$c$个彼此不同的类中心点$v_i$, $i$=1,2,…,$c$。

(2)计算样本点与中心点的距离$d_{ik}=\|x_k - v_i\|$，并且求出隶属度$u_{ik}$。

如果 $\forall k$ , $\forall i$ ,$d_{ik}\neq 0$，则

$$u_{ik} = d_{ik}^{\frac{2}{1-m}} \Big/ \left[\sum_{i=1}^{c} d_{ik}^{\frac{2}{1-m}}\right]$$

(3)如果存在$k$, $i$，使得$d_{ik}$=0，则令$u_{ik}$=1，且对$j\neq i$ $u_{jk}$=0；计算类中心点的坐标：

$$v_i = \sum_{k=1}^{n} (u_{ik})^m x_k \Big/ \left[\sum_{k=1}^{n} (u_{ik})^m\right]$$

(4)回到步骤(2)，直到

$$\left\|V^{(k+1)} - V^{k}\right\| < \mathrm{e}$$

其中，$V$ = [$v_1$,$v_2$,…,$v_c$]是类中心点的集合。

FCM 算法能快速、有效地对给定的数据集进行指定类数的聚类，以类中心点的位置和样本点对类的隶属度表示聚类结果，得出 $c$ 个模糊集合。

通常采用聚类有效性函数对聚类结果进行评价，包含度是一种常用的聚类有效性函数，描述了一个模糊集包含在另一个模糊集的程度。以聚类数$c$=2 为例，设聚类结果是两个模糊集合$F_1$和$F_2$，样本$x_k$对$F_1$和$F_2$的隶属度分别为$u_{1k}$和$u_{2k}$。定义$F_1$对$F_2$的包含度[5]为

$$\mathrm{sub}(F_1, F_2) = \sum_{k=1}^{n} \min(u_{1k}, u_{2k}) / \sum_{k=1}^{n} u_{1k} \tag{3}$$

FCM的聚类结果没有空类，所以，上式分母不为 0。类似可以定义$F_2$对$F_1$的包含度，最后给出聚类结果的有效性函数为[5]

$$Valid\text{-}F(U,c)=\max(\mathrm{sub}(F_1,F_2),\ \mathrm{sub}(F_2,F_1)) \tag{4}$$

式(4)以$F_1$对$F_2$的包含度和$F_2$对$F_1$的包含度中的较大者作为两个模糊集合$F_1$和$F_2$互相包含的程度度量。在聚类中，该函数值越小表示两个类别分开得越远，聚类效果越好。

## 4 实验设计与结果分析

### 4.1 实验方案设计

实验针对以下几个类进行：二尖瓣闭锁不全(Mitral incompentence)，二尖瓣狭窄(Mitral stenosis)，主动脉闭锁不全(Aortic incompence)，静脉曲张(Varicose veins)和糖尿性肾病(Diabetic Nephropathy)，其中，前 4 个类属于心血管类疾病，第 5 个类是内分泌系统疾病。样本网页来自 Medline 数据库，以类别名称作关键词，该词出现在网页中即认定和类别相关，再进一步通过阅读网页，剔除同时属于多个类别的网页。各类别的样本情况和预处理后结果如表 1 所示。

**表 1 实验类别基本情况**

| 类别名称 | 样本网页个数 | 频繁特征词数量 |
|---|---|---|
| 二尖瓣闭锁不全 | 288 | 942 |
| 二尖瓣狭窄 | 293 | 968 |
| 主动脉闭锁不全 | 242 | 841 |
| 静脉曲张 | 231 | 905 |
| 糖尿性肾病 | 314 | 1 798 |

实验研究任意两个类别在全局模型、局部模型和差异模型表示下的边界情况，将两个类别的样本混在一起，利用 FCM 算法进行 $c$=2 的聚类，利用包含度来评价聚类结果的有效性。

实验步骤如下：

(1)将网页转换成文本文件，按 3.1 节进行预处理，得出每个类的频繁词向量。

(2)任选两个类别，将样本混合起来，采用全局模型表示两个类中的每个文档，得到词项×文档矩阵$\boldsymbol{X}$，矩阵元素$x_{ij}$表示词项$t_i$在文档$d_j$中的出现次数。

(3)对矩阵$\boldsymbol{X}$进行SVD分解，设置$k$=20，取$\boldsymbol{U}$的前 20 维$\boldsymbol{U}_k$，然后将文档集合映射到低维空间表示为$\boldsymbol{X}'=\boldsymbol{U}^{T}{}_{k}\boldsymbol{X}$。

(4)对$X'$进行FCM聚类，按式(4)计算相应的包含度值[5]。

(5)将步骤(2)中的全局模型换成局部模型和差异模型重复该实验。

步骤(3)中的 SVD 分解是利用 Matlab 中的 SVD( )来完成的，其余代码使用 C 语言完成。

### 4.2 实验结果及分析

表 2 是对全局模型、局部模型和差异模型的对比描述。局部模型是用一个类别的特征向量来表示这两个类别混合到一起后文本集的特征向量($\boldsymbol{A}$ 向量是指用 $\boldsymbol{A}$ 的特征向量来表示文本集合 $\boldsymbol{A}\cup\boldsymbol{B}$)。全局模型的特征向量是 $\boldsymbol{A}$, $\boldsymbol{B}$ 特征向量的并集，而差异模型的特征向量是 $\boldsymbol{A}$, $\boldsymbol{B}$ 特征向量的对称差。通过表 2 中 3 种模型的包含度比较可以看出：用差异模型来表示文档时，类之间的包含度更低，两个类别的边界更清晰。通过统计发现：采用差异模型时的类间包含度平均约为全局模型时的 85%，约为局部模型时的 80%(按 $A$ 类向量计算)。

**表 2 不同模型表示下的聚类包含度比较**

| 类别/包含度 | | 差异模型 | 全局模型 | 局部模型 | |
|---|---|---|---|---|---|
| $A$ 类 | $B$ 类 | | | $\boldsymbol{A}$ 向量 | $\boldsymbol{B}$ 向量 |
| 主动脉闭锁不全 | 糖尿性肾病 | 0.575 407 | 0.661 739 | 0.758 599 | 0.755 267 |
| 主动脉闭锁不全 | 二尖瓣闭锁不全 | 0.980 027 | 0.996 279 | 0.990 824 | 0.989 191 |
| 主动脉闭锁不全 | 二尖瓣狭窄 | 0.634 628 | 0.895 548 | 0.894 961 | 0.890 452 |
| 主动脉闭锁不全 | 静脉曲张 | 0.647 109 | 0.743 331 | 0.785 942 | 0.839 902 |
| 糖尿性肾病 | 二尖瓣闭锁不全 | 0.591 232 | 0.659 152 | 0.749 092 | 0.739062 |
| 糖尿性肾病 | 二尖瓣狭窄 | 0.625 148 | 0.704 036 | 0.789 944 | 0.749 229 |
| 糖尿性肾病 | 静脉曲张 | 0.630 191 | 0.758 795 | 0.802 587 | 0.908 257 |
| 二尖瓣闭锁不全 | 二尖瓣狭窄 | 0.610 342 | 0.861 538 | 0.908 661 | 0.855 062 |
| 二尖瓣闭锁不全 | 静脉曲张 | 0.619 515 | 0.742 879 | 0.784 234 | 0.844 447 |
| 二尖瓣狭窄 | 静脉曲张 | 0.778 981 | 0.820 774 | 0.847 693 | 0.883 508 |

包含度越大说明两个类的包含程度越高，边界就越模糊。从二尖瓣闭锁不全和主动脉闭锁不全两类的包含度看出，这两个类是很相近的。实际上通过阅读网页发现，这两个类的网页讨论的内容基本上都涉及肺动脉回流、瓣膜性回流、肥原型心肌、心室间隔缺损、左主动脉、急性风湿等内容，具有很大的相似性。糖尿性肾病类与其他类别的包含度都很低，事实上这个类与其他类别的疾病分别属于两个不同种类的疾病，糖尿性肾病是属于内分泌系统方面的疾病，而其余 4 类是属于心血管方面的疾病，它们本身的相似性就很低，实验也验证了这一点。

## 5 结束语

随着 Internet 和 Web 挖掘技术的快速发展，数据挖掘技术不再局限于结构化数据的挖掘。如何从 Web 文本中发现隐含的、有意义的信息，已成为数据挖掘的重要目标。在本文实验中，每一类的特征向量都是由文档频率 > 4 的特征词组成的。实验中将两个类的数据混合在一起进行聚类，并计算类间的包含度，从而比较不同的潜在语义模型得到的向量对聚类的影响。实验结果表明，用潜在语义差异模型得到的特征向量使类别之间的边界更清晰，因此，可以认为用潜在语义差异模型来表示文本的特征效果更好。下一步将此模型用于 Web 文本分类，并与全局模型和局部模型的改进模型进行比较。

### 参考文献


[1] Kazuo H, Matsumoto Y. Information Extraction from Medline Abstracts of Clinical Trials[J]. Technical Report of IEICE AI, 2004, 42(12): 45-49.

[2] 孙建涛. Web 挖掘中的降维和分类方法研究[D]. 北京: 清华大学, 2005.

[3] Berry M W, Dumais S T, O'brien G W. Using Linear Algebra for Intelligent Information Retrieval[J]. SIAM Review, 1995, 37(4): 573-595.

[4] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[5] 范九伦, 吴成茂. 用于聚类有效性判定的包含度公式[J]. 模糊系统与数学, 2002, 16(1): 80-86.


（上接第 63 页）

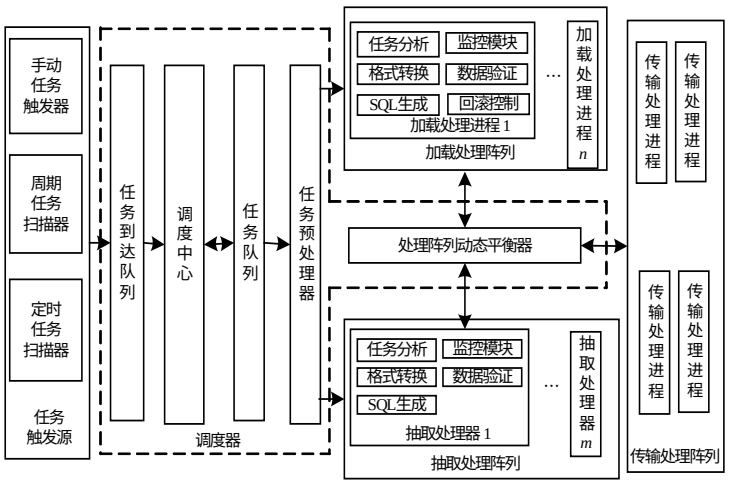


**图 5 交换引擎结构**

## 5 结束语

以修订的信息结构体系[1]为标准，该平台方便了组织、干部系统异构数据库间的数据交换。目前平台主要考虑的是关系数据库间的信息交换，以后可以考虑：(1)其他格式化文档与数据库间的交换；(2)在双方数据库都符合信息结构体系的情况下的自动数据交换。

### 参考文献


[1] 中共中央组织部. 全国组织干部人事管理信息系统信息结构体系[Z]. 2006.

[2] Fuxman A, Phokion G. Peer Data Exchange[C]//Proc. of TODS'06. [S. l.]: ACM Press, 2006: 1454-1498.

[3] Lucian P. Translation Web Data[C]//Proc. of VLDB'02. Hong Kong, China: [s. n.], 2002: 598-609.

[4] Arenas M, Libkin L. XML Data Exchange: Consistency and Query Answering[C]//Proceedings of the 24th ACM Sigmod-sigact-sigart Symposium on Principles of Database Systems. New York, USA: ACM Press, 2005.

[5] Audsley N C. Deadline Monotonic Scheduling[D]. Charlottesville, Canada: Department of Computer Science, University of Virginia, 1990.