文章编号:1005－0523(2004)05－0059－04

WEB 页面的模糊聚类

钟茂生

(华东交通大学 信息工程学院 江西 南昌 330013)

摘要:当前 Internet 上信息浩如烟海,人们很容易从 Internet 上获得大量的信息资料,然而如何对这些信息资料进行有效的管理和分类是摆在人们面前的一个不可回避而又很有意义的问题.本文通过采用模糊集合理论和模糊聚类技术,对 Internet 上的 WEB 页面进行聚类从而将 WEB 页面自动分类.实验表明,对 WEB 页面进行模糊聚类是一种有效的方法.

关 键 词:WEB 页面;模糊集合;模糊关系;相似度;模糊聚类

中图分类号:TP391　　**文献标识码**:A

1 引 言

随着科学技术的迅猛发展,特别是 Internet 的迅猛发展,各种信息情报激增,WEB 信息浩如烟海,因此人们可以通过 Internet 很快地得到大量的资料.然而如何对所得到的资料进行有效的管理是摆在人们面前的一个不可回避而又很有意义的问题.对资料进行管理的一个常见的方法就是将他们进行系统的分类.显然,用人工对文本材料进行分类的过程是通读所有文章,然后将其归类保存.当然这需要许多具有丰富经验和专门知识的分类人员做大量的工作,这个过程具有周期长、费用高、效率低的缺陷,在这个信息爆炸的今天很难满足实际需要,如何用计算机对 WEB 页面或文本材料进行自动分类成了许多人的研究方向.

目前,关于 WEB 页面或文本材料的自动分类,人们提出了许多方法,如文献[3]中提出用基于粗糙集的关键词维数约简技术(RSDR)对文本进行分类;文献[4]中提出用 NaiveBayes 方法协调分类 Web 网页;文献[5][6]也提出了一些文本分类的算法或模型,等等.这些分类方法总体上可以分为两种类型[6]:①基于外延方法的分类,即不关心文本的语义,根据文本的外在特征进行分类.如基于向量空间模型(VSM)的方法;②基于语义的分类方法,即采用全部或部分理解文本的语义进行分类,如基于词的归类,基于知识的归类,基于概念的归类.

当前,模糊集合和模糊聚类问题在人工智能和数据挖掘得到了大量的研究,关于模糊聚类的研究工作可分为两种:1)通过模糊方法得到模糊结果,每一个对象可以不同的隶属度从属于若干个类;2)通过模糊的方法得到确定的结果,如文献[7]介绍了用模糊聚类技术对 Internet 中的客户群体进行聚类(其聚类方法是直接用模糊相似矩阵进行,但其聚类结果有公共元素,即一个元素可从属于多个聚类).本文提出用模糊集理论和模糊聚类技术对 WEB 页面进行分类,首先建立 Web 页面模糊集,用余弦幅度相似性度量规则构造相应的模糊相似矩阵,然后根据模糊相似矩阵进行聚类,实验表明该方法能够进行有效聚类.

2 模糊集理论简介

模糊集理论是 Zadeh 博士在 1965 年提出的[1],

收稿日期:2003－10－16

作者简介:钟茂生(1973－),男,江西兴国县人,华东交通大学讲师.

他用集合隶属度的概念来处理不确定或模糊性信息,他指出,经典集合中的对象具有精确的隶属度,而模糊集合中的对象具有近似的隶属度.在经典集合中,在论域 U 上的元素 x 要么是某个清晰集合 A 的元素,要么不是;而模糊集合中,论域 U 上元素 x 用隶属度函数 $\mu_A(x)$来表示, $\mu_A(x)$是单位区间上衡量元素 x 属于模糊集合 A 的一个值.

用 U 表示论域,$U=\{u_1,u_2,\cdots,u_n\}$,假设 A 是 U 上的一个模糊集,那么,A 可以表示为:$A=\{(u_1,\mu_A(u_1)),(u_2,\mu_A(u_2)),\cdots,(u_i,\mu_A(u_i)),\cdots,(u_n,\mu_A(u_n))\}$

其中 $\mu_A:U\to[0,1]$是模糊集 A 的隶属函数;$\mu_A(u_i)$则为 u_i 在 A 中的隶属度.

设模糊集 A,B 的隶属函数分别为 μ_A,μ_B.对于模糊集 A,B 可作下列基本运算:

1) 交集 :$\mu_{A\cap B}(u_i)=\mathrm{Min}(\mu_A(u_i),\mu_B(u_i)),u_i\in U$.

2) 并集 :$\mu_{A\cup B}(u_i)=\mathrm{Max}(\mu_A(u_i),\mu_B(u_i)),u_i\in U$.

3) 补集 :$\mu_A(u_i)=1-\mu_A(u_i),u_i\in U$.

模糊集合与清晰集合有相同的性质,同样满足交换律、结合律、分配律、互补律等.此外,可以将清晰集合的隶属度值看作是[0,1]区间上的一个子集,清晰集合可以认为是模糊集合上的一个特例.

与清晰关系一样,模糊关系也是通过两个论域上的笛卡尔积把一个叫 X 的论域中的元素映射到另一个叫 Y 的论域上去,不过这两个论域上的序偶间的关系"强度"不是用特征函数来测量,而是用隶属度函数在单位区间[0,1]上的不同的值来表示其关系"强度",因此模糊关系 $\Re$ 是笛卡尔积 $X\times Y$ 到区间[0,1]上的映射,其映射的强度可用从两个论域或 $\mu_{\Re}(x,y)$的序偶关系的隶属函数来表示.

设 $\Re_1$ 和 $\Re_2$ 是笛卡尔空间 $X\times Y$ 上的模糊关系,下列运算可为各种集合运算提供隶属值:

并:$\mu_{\Re_1\cup\Re_2}(x,y)=\mathrm{Max}(\mu_{\Re_1}(x,y),\mu_{\Re_2}(x,y))$

交:$\mu_{\Re_1\cup\Re_2}(x,y)=\mathrm{Min}(\mu_{\Re_1}(x,y),\mu_{\Re_2}(x,y))$

补:$\mu_{\overline{\Re_1}}(x,y)=1-\mu_{\Re_1}(x,y)$

包含:$\Re_1\subset\Re_2\Rightarrow\mu_{\Re_1}(x,y)\le\mu_{\Re_2}(x,y)$

假定 $\Re$ 是 $X\times X$ 上的模糊关系,则由模糊关系 $\Re$ 构造的模糊矩阵 $M\in\xi(X\times X)$具有如下性质:

1)自反性:$x\Re x;x\in X$;(或 $\mu_{\Re}(x,x)=1$)

2)对称性:$x\Re y\Rightarrow y\Re x,x,y\in X$.(或 $\mu_{\Re}(x,y)=\mu_{\Re}(y,x)$)

上述模糊关系 $\Re$ 又称模糊相似关系,其构造的模糊矩阵 M 一般不满足传递律 ,是一种相似矩阵.通过计算其传递闭包(即 $R^{n-1}=R\circ R\circ\cdots\circ R$),可将其变为满足传递律的模糊等价矩阵,而满足传递性的模糊关系 $\Re$ 则为模糊等价关系 $\Re'$,即 $x\Re'y\wedge y\Re'z\Rightarrow x\Re'z.x,y,z\in X$.(或 $\mu_{\Re'}(x,y)=\lambda_1\wedge\mu_{\Re'}(y,z)=\lambda_2)\Rightarrow\mu_{\Re'}(x,z)=\lambda$,其中 $\lambda\ge\mathrm{Min}[\lambda_1,\lambda_2]$).

3 模糊聚类方法

聚类过程[1]就是确定一个由 n 个数据样本组成的数据空间 X 中的子类聚类的数目 c,并将该 X 划分成 c 个聚类的过程($2\le c\le n$).实际上,当 $c=1$ 时表明数据中存在聚类的假设是不存在的,而 $c=n$ 时说明每一个样本所在的"聚类"是由该样本自身组成的.有两种将数据划分成 c 个聚类的方法:硬(或清晰)和软(或模糊)的方法,本文主要讨论模糊聚类方法,关于硬(清晰)聚类方法请参考文献[2].

模糊聚类的方法很多,本文利用模糊等价关系将给定对象划分为一些等价类,依此进行聚类.设 X 为待分类的 n 个对象的集合,$X=\{X_1,X_2,\cdots,X_i,\cdots X_n\}$.每一对象 X_i 由 m 个特征定义,即用一组特征数据 $(x_{i1},x_{i2},\cdots,x_{im})$表示,因此 X_i 是一个有 m 个元素或 m 个特征的 m 维向量,在空间上,每个 X_i 就是 m 维特征空间中的一个点,而 X 则是空间中含有 n 个元素的一个点集,模糊聚类方法的基本步骤是:

①采用某种相似性度量方法如切比雪夫距离、海明距离、欧氏距离以及相似系数法、贴近度法等(本文用余弦幅度相似性度量,见(1)式)确定对象 X_i 与对象 X_j 之间的相似程度,S_{ij}^{μ},依此建立模糊相似矩阵 $M_{n\times n}^{\mu}$.

$$S_{ij}^{\mu}=\frac{\sum_{k=1}^{m}\chi_{ik}\times\chi_{jk}}{\sqrt{(\sum_{k=1}^{m}\chi_{ik}^2)\times(\sum_{k=1}^{m}\chi_{jk}^2)}}\qquad(1)$$

②通过对模糊相似矩阵的至多 n-1 次复合,计算其传递闭包,得到模糊等价矩阵 $M_{n\times n}^{\mu}$.因为按照上述方法构造的模糊相似矩阵是自反的和对称的,一般不具有传递关系,但通过至多 n-1 次的复合后,模糊相似矩阵 $M_{n\times n}^{\mu}$后可重新组合成一个模糊等价矩阵 $M_{n\times n}^{'\mu}$.在模糊等价矩阵 $M_{n\times n}^{'\mu}$中,其传递关系要满足的条件是:对所有 $\mu_{\Re'}(x_i,y_j)=\lambda_1\wedge\mu_{\Re'}(y_j,x_k)=\lambda_2\Rightarrow\mu_{\Re'}(x_i,x_k)=\lambda$,其中 $\lambda\ge\mathrm{Min}[\lambda_1,\lambda_2]$.

③根据设定的 λ 值(λ 分割), 对模糊等价矩阵

$M_{n}^{''\mu}{}_{\times n}$进行 λ 分割得到 λ－截矩阵 $M_{n}^{''\mu}{}_{\times n}$,由此 λ－截矩阵将对象集合划分为一些等价类,这些等价类即为聚类的结果.

例:假定有 5 个数据点 X_1, X_2, X_3, X_4, X_5,其构建的模糊相似矩阵 $M_{5\times5}$如下所示:

$$M_{5\times5}=\begin{bmatrix}1 & 0.8 & 0 & 0.1 & 0.2\\ 0.8 & 1 & 0.4 & 0 & 0.9\\ 0 & 0 & 1 & 0 & 0\\ 0.1 & 0 & 0 & 1 & 0.5\\ 0.2 & 0.9 & 0 & 0.5 & 1\end{bmatrix}\Rightarrow$$

$$M_{5\times5}^{4}=\begin{bmatrix}1 & 0.8 & 0.4 & 0.5 & 0.8\\ 0.8 & 1 & 0.4 & 0.5 & 0.9\\ 0.4 & 0.4 & 1 & 0.4 & 0.4\\ 0.5 & 0.5 & 1 & 0.5 & \\ 0.8 & 0.9 & 0.4 & 0.5 & 1\end{bmatrix}\Rightarrow$$

$$M_{5\times5}^{'}=\begin{bmatrix}1 & 1 & 0 & 0 & 1\\ 1 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0\\ 1 & 1 & 0 & 0 & 1\end{bmatrix}$$

矩阵 $M_{5\times5}$是自反的和对称的,但没有传递性,如:$\mu_{\Re}(X_1,X_2)=0.8$, $\mu_{\Re}(X_2,X_5)=0.9\geq0.8$,但 $\mu_{\Re}(X_1,X_5)=0.2\leq\mathrm{Min}(0.8,0.9)$.该矩阵经过 4 次复合后变为模糊等价矩阵 $M_{5\times5}^{4}$.

假定对该模糊等价矩阵 $M_{5\times5}^{4}$进行 λ 分割(λ=0.8)则可得到 λ－截矩阵 $M_{5\times5}^{'}$.经过等价类划分后即可得到聚类结果$\{X_1,X_2,X_5\}$, $\{X_3\}$, $\{X_4\}$.

4　WEB 页面的模糊聚类

4.1　Web 页面聚类步骤

根据上述模糊聚类算法,假定要对 n 个 Web 页面进行聚类,其基本步骤是:

①将每个 Web 页面看作是由相互独立的词条组$(T_1,T_2,\ldots,T_m)$构成,对于每一词条 T_i,都根据其在文档中的重要程度赋以一定的权值 W_i,并将 T_1, T_2, …, T_m 看成一个 m 维坐标系中的坐标轴,W_1, W_2, … W_m,为对应的坐标值,因此,每个 Web 页面将看作 m 维空间中的一个点.用余弦幅度相似性度量方法(见(1)式)计算 n 个 Web 页面内每两个页面之间的相似度 S_{ij},由此构造相似矩阵 $M_{n\times n}$.上述 Web 页面词条的特征抽取方法可参见文献[8],词条的权值计算方法可参见文献[3]中基于粗糙集的属性重要程度(degree of dependency of attributes)计算方法或文献[6]中的关键词集的抽取方法.

②因为按照步骤①构造的相似矩阵 $M_{n\times m}$具有自反性和对称性,但不一定具有传递性,为了使相似矩阵 $M_{n\times m}$具有传递性,可根据上述模糊聚类算法对其进行不超过 n－1 次的复合即可得到等价矩阵 $M_{n\times m}^{'}$.

③设定一 λ 值,对等价矩阵 $M_{n\times n}^{'}$进行 λ 分割和进行等价类划分后,即可得到等价类,此等价类即为相关 Web 页面的聚类结果.

4.2　WEB 页面模糊聚类实验

根据上述方法,我们随机从 Internet 下载了 100 个 Web 页面进行聚类实验,这些 Web 页面根据其标题从直观上看可以分为搜索引擎、数据挖掘、算法研究、GIS(地理信息系统)、Rough set 理论等类别,在实验过程中,我们采用了 75957 个中文词汇和 GIS、Rough set、Web 三个英文词汇(因为这 100 个 Web 页面标题中含有这三个单词/词组),通过使用这些词汇对这些 Web 页面的标题进行分词、去除常用词、特征抽取和赋权值等处理后,用(1)式计算任意两个页面之间的相似度,得到模糊相似矩阵 $M_{100\times100}$,然后计算传递闭包,设定 λ=0.82,得到 λ－截矩阵 $M_{100\times100}^{'}$,最后得到 22 个聚类结果.通过分析这些结果发现,有 78 个 Web 页面聚在上述 5 个类中,另外 22 个 Web 页面中,X_7, X_9, X_{10}, X_{18}, X_{19}, X_{31}, X_{40}, X_{X53}, X_{60}, X_{69}, X_{71}, X_{97},自身成为一类,而$\{X_3, X_4\}$, $\{X_{16}, X_{67}\}$, $\{X_5, X_{65}\}$, $\{X_{85}, X_{86}\}$, $\{X_{98}, X_{99}\}$各成一个类,如类$\{X_3, X_4\}$为两种杂志文章的“总目录”介绍,类$\{X_{16}, X_{67}\}$为两个决策支持系统的介绍等等.这个聚类实验表明,只要中文词汇较全、对 Web 页面特征抽取准确、相似性度量科学,用这种模糊聚类的方法对 Web 页面聚类还是很有效的.

我们在实验中发现的问题主要是:在对 Web 页面文档(实际上对其它所有的文档也是一样)进行聚类之前,如果能对所有汉字或词组从语义上进行有效的分类,则对 Web 页面聚类的效果较好,因为这能减少文档向量的维数,从而相应减少计算量,而且这种聚类可以把文档和其语义结合起来进行.而要对汉字或词组进行有效的分类,可对一定数量的文档进行训练来获得结果,这种方法具体可参见相关文献.

5　结束语

关于用模糊聚类技术对 Web 页面进行聚类,本

文讨论了用模糊等价关系进行聚类的方法，但模糊聚类还有其它很多方法，其中非常流行的一种就是称为模糊 c－均值的算法，该方法的实质就是固定将样本数据聚成 c 个类，如果用该方法对 Web 页面进行聚类，遇到的问题就是：很难人为确定 c 的值；要反复计算分区矩阵 $U^{(\gamma)}$，直到 $\| U^{(\gamma)} - U^{(\gamma-1)} \| \leq \varepsilon$（$\varepsilon$ 为预定的精度），计算量大. 本文用模糊等价关系的方法，计算量要小，能达到预定的效果. 此外，如果用模糊等价关系先进行聚类，可大致确定聚类的数目 c，然后再用模糊 c－均值的算法再进行聚类，即将两种方法结合起来，这是一种值得考虑的聚类方法.

参考文献：

[1] Timothy J. Ross 著，钱同惠，沈其聪，葛晓滨，等译. 模糊逻辑及其工程应用（Fuzzy Logic With Engineering Applications）[M]. 北京：电子工业出版社，2001.

[2] Jiawei Han & Micheline Kamber. Data mining：Concepts and Techniques[M]. Academic Press，2000.

[3] Alexios Chouchoulas & Qiang Shen. Rough Set－Aided Keyword Reduction For Text Categorization[J]. Application Artificial intelligence，2001，8：857－861.

[4] 范焱，郑诚，王清毅，蔡庆生，刘洁. 用 Naive Bayes 方法协调分类 Web 网页[J]. 软件学报，2001.9：1386－1389.

[5] 徐德智，阳绿云. 中文网叶自动分类研究[J]. 计算机工程与科学，2001，6：33－35.

[6] 苏伟峰，李绍滋，李堂秋. 一个基于概念的中文文本分类模型[J]. 计算机工程与应用，2002，6：193－195.

[7] 宋擒豹，沈钧毅. Web 页面和客户群体的模糊聚类算法[J]. 小型微型计算机系统，2001，2：44－49.

[8] 李凡，鲁明羽，陆玉昌. 关于文本特征抽取新方法的研究[J]. 清华大学学报（自然科学版）. 2001，41，(3)：98－101.

Fuzzy Clustering of WEB Page

ZHONG Mao-Sheng

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: Nowdays, there are much information in Internet, but how to manage and classify it is an inevitable and meaningful thing This paper described a method in which WEB pages are clustered or auto－classified by using FUZZY SET theory and FUZZY CLUSTERING technique. Experiment of the WEB page's FUZZY CLUSTERING indicates that FUZZY CLUSTERING is effective method for WEB page.

Key words: WEB page; fuzzy set; fuzzy relation; similarity; fuzzy clustering