基于不可分辨关系的文本自动聚类①

周 勇

(四川广播电视大学 高职学院, 成都 610073)

摘 要: 研究了文本对象在不可分辨关系下的自动聚类方法. 在自动聚类过程中, 首先把文本集转化为让机器可以处理的布尔文本信息系统; 其次在信息系统上定义对象间的不可分辨关系, 提出利用不可分辨关系进行聚类的理论基础; 然后对算法进行描述, 并用实验进行验证; 最后分析该算法的时间复杂度和缺点, 并提出具体的改进措施. 基于不可分辨关系的文本自动聚类算法具有理论基础和较好的实验效果表明该方法具有较好的应用性.

关键词: 文本信息系统; 不可分辨关系; 文本自动聚类

Text Automatic Clustering Based on Indiscernibility Relation

ZHOU Yong

(School of Higher Vocation, Sichuan Radio and TV University, Chengdu 610073, China)

Abstract: This paper studied the automatic clustering method under the Indiscernibility relation of the text objects. In the clustering process, the text sets were converted to the Boolean text information system that the machine may process; secondly the Indiscernibility relation was defined in information systems, and the Indiscernibility relation clustering theory was proposed; then the algorithm was described, which was proved by experiment; Analyzing the time complexity and disadvantages of the algorithm, gives the concrete improvement measures. Based on Indiscernibility relation automatic text clustering algorithm has a theoretical foundation and good experimental results show that this method has better application.

Key words: text information system; indiscernibility relation; text automatic clustering

聚类是把多维空间内的数据集合分成多个有意义的子群或者类的过程, 使同类中的样本相似度尽可能大, 不同类的样本相似度尽可能小. 文本聚类就是把文本训练集分成子群或者类的过程, 它是一种无监督的机器学习方法.

目前, 比较成熟的文本聚类方法是文本向量空间聚类法, 一般步骤为文本分词、特征词选取、文本向量化、设计文本间的相似性函数和文本自动聚类. 设计文本对象间的相似性函数是文本聚类的重要环节, 不同相似性函数决定不同聚类结果[1,2]. 本文的文本自动聚类算法是从文本自身特性出发, 没有加入人为定义的文本相似性函数, 使得聚类结果更具客观性.

1 文本集信息系统化

定义 1.1[3]: 称四元组 $S=<D,T,V,f>$ 为文本信息系统, 其中 $D=\{d_1,d_2...d_n\}$ 是文本集, $T=\{t_1,t_2...t_m\}$ 是全体特征词的集合, $V=\bigcup_{t\in T}V_t$, V_t 是特征词 t 的值域, $f:D\times T\to V$ 是信息函数, 满足 $f(d,t)\in V$, $d\in D$, $t\in T$.

表 1 布尔文本信息系统

D	t_1	t_2	...	t_m
d_1	0	1	...	1
d_2	1	0	...	0
...	...	...	...	...
d_n	1	1	...	1

① 收稿时间:2012-05-11;收到修改稿时间:2012-06-12

根据特征词属性值表示的意义不同, 把文本信息系统分成不同的种类. 如果特征词的属性值表示特征词在该文本中是否出现, 那么称为布尔文本信息系统. 如果特征词的属性值为特征词的 $tf-idf$ 值, 那么称为 $tf-idf$ 文本信息系统. 如果特征词的属性值表示特征词在该文本中出现的次数, 那么称为词频文本信息系统. 在布尔文本信息系统中, 特征词的取值一般为 0 和 1, 在 $tf-idf$ 文本信息系统中, 特征词的属性值取值范围为 $[0.+\infty)$, 在词频文本信息系统中, 特征词的属性值的取值范围为自然数集. 当然, 布尔文本信息系统为词频文本信息系统的特殊情况, 表 2-1 是布尔文本信息系统.

基于不可分辨关系的文本自动聚类首先把文本集转化成文本信息系统, 其转化步骤为文本分词、特征词选取、确定特征词的属性值; 本文的文本自动聚类要求把文本集转化成布尔文本信息系统.

算法 1 文本集信息系统化

输入: 文本集

输出: 布尔文本信息系统 $S=<D,T,V,f>$;

Step1: 应用中科院分词系统对文本进行分词;

Step2: 选择特征词, 并合并所有特征词形成特征词集;

Step3: 搜索每个特征词在文本是否出现, 并赋予属性值为 0 或者 1;

Step4: 输出布尔文本信息系统.

2 不可分辨关系

定义 2.1. 在文本信息系统 $S=<D,T,V,f>$ 中, 任意 $t\in T$, 称二元关系:

$$IND(t)=\{(x,y,)\in D\times D \mid f(x,t)=f(y,t)\}$$

为 t 上的不可分辨关系, 记为 $IND(t)$.

定义 2.2 在文本信息系统 $S=<D,T,V,f>$ 中, 给定 $T_0(T_0\subset T)$, 称二元关系:

$$IND(T_0)=\{(x,y,)\in D\times D \mid \forall t\in T_0, f(x,t)=f(y,t)\}$$

为 T_0 上的不可分辨关系族, 记为 $IND(T_0)$.

$IND(t)$ 是文本信息系统 $S=<D,T,V,f>$ 的不可分辨关系, 如果 $(x,y)\in IND(t)$, 那么记作 $xIND(t)y$.

定理 2.1. 给定文本信息系统 $S=<D,T,V,f>$, 任意 $t\in T$, 不可分辨关系 $IND(t)$ 是 D 上的等价关系.

证明: 只需证明可分辨关系 $IND(t)$ 具有自反性、对称性和传递性即可.

(1) 自反性 由不可分辨关系可得对于任意 $x\in D$ 都有 $(x,x)\in IND(t)$;

(2) 对称性 任意 $(x,y)\in IND(t)$, 因为有 $f(x,t)=f(y,t)$, 所以一定有 $(y,x)\in IND(t)$;

(3) 传递性 任意 $(x,y)、(y,z)\in IND(t)$, 因为 $f(x,t)=f(y,t)$ 和 $f(y,t)=f(z,t)$ 一定有 $f(x,t)=f(z,t)$, 所 $(x,z)\in IND(t)$;

所以不可分辨关系 $IND(t)$ 是 D 上的等价关系.

定理 2.2. 给定文本信息系统 $S=<D,T,V,f>$, 给定 $T_0(T_0\subset T)$, 不可分辨关系族 $IND(T_0)$ 是 D 上的等价关系.

定理 2.3. 文本信息系统 $S=<D,T,V,f>$ 上任意不可分辨关系 $IND(t)$ 确定 D 上的一个划分, 记作 $D/IND(t)$, 即有:

$$D/IND(t)=\{D_1,D_2..D_l \mid D_i\subseteq D, D_j\subseteq D, \forall i\neq j,$$
$$(i,j=1,2..l)D_i\cap D_j=\varnothing, \bigcup_{i=1}^{l} D_i=D\}$$

定理2.4. 文本信息系统 $S=<D,T,V,f>$ 上任意不可分辨关系 $IND(T_0)$ 确定 D 上的一个划分, 记作 $D/IND(T_0)$, 即有:

$$D/IND(T_0)=\{D_1,D_2..D_l \mid D_i\subseteq D, D_j\subseteq D, \forall i\neq j,$$
$$(i,j=1,2..l)D_i\cap D_j=\varnothing, \bigcup_{i=1}^{l} D_i=D\}$$

根据定理 2.4, 在 D 上定义的不可分辨关系可以把 D 划分成互不相交的子集, 而这个划分具有的特点为: (1)相同子集中任意两个文本的特征词取值都全部相同; (2)不同子集中任意两个文本至少有一个特征词的取值不相同. 这种划分的特点与聚类两大假设完全吻合.

定义 2.3. $IND(t)$ 是文本信息系统 $S=<D,T,V,f>$ 上的不可分辨关系, 任意对象 $x_i\in D$, 称 D 的子集 $[x_i]_{IND(t)}=\{x_j \mid x_j\in D, 且\ x_i IND(t) x_j\}$ 为对象 x_i 在不可分辨关系 $IND(t)$ 下的生成等价类.

定理 2.5. 文本信息系统 $S=<D,T,V,f>$, $IND(t)$ 是上 S 的不可分辨关系, 由对象 x 在不可分辨关系 $IND(t)$ 下生成的等价类 $[x]_{IND(t)}$ 具有以下性质:

(1)若 $x_j\in[x]_{IND(t)}$, 则 $[x_j]_{INT(t)}=[x]_{IND(t)}$;

(2)若 $x_j\notin[x]_{IND(t)}$, 则 $[x_j]_{IND(t)}\cap[x]_{IND(t)}=\varnothing$.

3 算法描述及分析

定理 2.5. 给出了基于不可分辨关系的文本自动聚类方法, 其核心步骤为搜索每个对象的生成等价类, 最后输出所有等价类.

算法 2. 基于不可分辨关系的文本自动聚类

输入: 布尔文本信息系统 $S=<D,T,V,f>$

输出: 文本聚类集 $\{D_1,D_2...D_m\}$

Step1: 搜索对象 d_1 的生成等价类 $[d_1]_{IND(T)}$;

Step2: 搜索对象 d_i 的生成等价类 $[d_i]_{IND(T)}$.

判断 d_i 是否属于 $[d_1]_{IND(T)}$, $[d_2]_{IND(T)}\cdots[d_{i-1}]_{IND(T)}$ 中某个生成等价类.

如果 d_i 属于 $[d_1]_{IND(T)},[d_2]_{IND(T)}...[d_{i-1}]_{IND(T)}$ 中某个对象生成的等价类, 那么 $i=i+1$ 转到 Step2;

如果 d_i 不属于 $[d_1]_{IND(T)},[d_2]_{IND(T)}...[d_{i-1}]_{IND(T)}$ 中任意一个对象的生成等价类, 那么搜索该对象生成的等价类 $[d_i]_{IND(T)}$. 直到搜索完所有的对象为止.

Step3: 输出所有对象的生成等价类.

4 算法和实验结果分析

算法的最坏情况为每个对象的生成等价类中的元素只有它自身, 此时自动聚类的时间最长. 当文本集的文本对象数为 n 时, 时间复杂度为 $O(n^2)$.

为了测试该文本聚类算法的有效性, 选择了 100 篇不同类型的文章, 利用中科院的分词系统, 对每个文本进行分词, 利用 $tf-idf$ 阈值法选择每个文本的特征词, 把该文本集转化成布尔文本信息系统, 最后利用不可分辨关系进行文本自动聚类, 其聚类结果如表 2 所示.

表 2 不可分辨关系聚类比较分析

文本类型	数量	聚类结果	正确分类比例
金融类	10	8	80%
体育类	20	15	75%
娱乐类	28	20	71%
旅游类	20	14	60%
IT 类	22	23	
其他 1		3	
其他 2		2	
其他 3		7	
其他 4		5	
其他 5		3	

从上表可以看出, 基于不可分辨关系的文本聚类方法具有较高的正确聚类比例, 但是出现分类比较细, 超出本身所具有类的数量, 本身有 5 类, 最后聚成 10 类, 出现这种情况主要有以下方面的原因:

(1) 人为聚类本身具有一定的局限性, 即这 100 篇文章本身就不只 5 类;

(2) 不可分辨关系苛刻的条件决定聚类结果较细, 可能出现某个文档独自成类的情况.

5 展望与改进

基于不可分辨关系的文本自动聚类算法是从文本自身的特点出发, 没有加入人为刻画文本相似性的因素, 因而聚类结果具有唯一性, 但是该方法却不可避免的出现聚类结果过于细化的现象. 为了避免这种弊端, 可以采取以下两种方式进行改进:

(1) 选择少量具有 “代表性”的特征词. 在文本转化所文本信息系统时, 每个特征词的重要性不尽相同. 在聚类时, 只选择相对重要的特征词, 这样减少特征词的数量, 可以避免出现聚类较细的现象.

(2) 放松不可分辨关系中特征词属性值 “相等”的条件. 在不可分辨关系定义中, 要求每个特征词的值都必须相等, 在聚类过程中还要求每个特征词的属性值都得相等. 这是产生聚类过细的根本原因, 因而可以放松 “相等”为 “近似”, 特别是在文本信息系统中, 只要重要的特征词的属性值有大于某个值就认为是对象在该特征词上不可分辨.

参考文献

1 吴启明,易云飞.文本聚类综述.河池学院学报,2008,28(2).

2 高茂庭.文本聚类分析若干问题研究[博士学位论文].天津: 天津大学,2006.13–14.

3 张文修,吴伟志,梁吉业,李德玉.粗糙集理论与方法.北京:科学出版社,2001.1–25.