◎数据库、信号与信息处理◎

一种基于 GN 算法的文本概念聚类新方法

安 娜,谢福鼎,张 永,刘绍海
AN Na,XIE Fu-ding,ZHANG Yong,LIU Shao-hai

辽宁师范大学 计算机与信息技术学院,辽宁 大连 116029
Department of Computer Science,Liaoning Normal University,Dalian,Liaoning 116029,China
E-mail:anan983@163.com

AN Na,XIE Fu-ding,ZHANG Yong,et al. New method for text concept clustering based on GN algorithm. Computer Engineering and Applications,2008,44(14):142-144.

Abstract: Text clustering is a basic and important topic in text mining. This paper presents a new text clustering method which takes the advantages of concept lattice and complex network. The algorithm firstly computes the weights of the key words and processes the problem of decreasing dimension,and then the formal context is constructed in terms of key words which have the proper weight. Secondly,the similarities between concepts are computed by using of the formula proposed in this paper. Text concept clustering can be done by the construction of GN network and application of GN algorithms. At last,the experiment shows the validity of this method.
Key words: complex networks;GN algorithm;text clustering;concept lattices

摘 要:文本聚类是当前文本信息挖掘的基础和研究的重点。给出一种新的文本聚类方法,它将概念格和复杂网络有机地结合起来,以达到更优的聚类效果。首先计算关键词特征权值并对特征向量进行降维处理,然后根据关键词权值大小映射到形式背景中,通过本文所给出的新的相似度公式,计算出形式背景中概念相似度的大小,从而构造 GN 网络并应用 GN 算法进行文本概念聚类。最后通过实例,验证了方法的可行性。
关键词:复杂网络;GN 算法;文本聚类;概念格
DOI:10.3778/j.issn.1002-8331.2008.14.039 **文章编号**:1002-8331(2008)14-0142-03 **文献标识码**:A **中图分类号**:TP391

1 引言

随着 Internet 的普及和信息量的激增,文本聚类已成为当前数据挖掘的重要研究方向之一。文本聚类方法较多,主要有划分的方法(以 k-means 为代表)[1]、凝聚的层次聚类方法[2]、基于密度的聚类方法[3]、基于 SOM 神经网络方法[4]等。各种方法有不同的特点,前两种方法是经典聚类算法在文本聚类方面的应用,通常层次聚类法的聚类效果比平面划分法好,但前者运行时间比后者高;基于密度的聚类方法能够处理任意形状和大小的簇;SOM 神经网络法抗噪声干扰性较强,它们在不同的领域都分别得到了成功的应用。

形式概念分析是从对象数据表里抽取信息的自然聚类分析方法。文本进行预处理后通过形式概念分析中的算法将文本集生成了文本概念集,这种由形式背景生成概念的过程实质上是一个概念聚类的过程。寻找复杂网络[5]中的社团结构,本质上也是一种聚类或者分类的过程。即把一个数据集分成或者聚成若干称为集簇的子集,每个簇中点间具有较大的相似性,而簇之间的点具有较小的相似性。

基于形式概念分析和复杂网络,本文提出了一种新的文本聚类方法,该算法将形式概念分析和复杂网络有机的结合,得到聚类结果。将社团结构的思想应用到无监督的文本概念聚类中。研究表明,本文在空间向量模型的基础上,将 GN 算法和形式概念分析理论相结合,为文本聚类提供了一个新的思路和方法。

2 预备知识

2.1 GN 算法

Girvan 和 Newnan[6]于 2001 年提出一个基于边介数的社团发现算法,该算法是一种分裂方法,依据边介数把不属于任何社团的边逐步删除。边介数定义为网络中经过每条边的最短路径数目,即所有最短路径通过该边的次数之和为该边的

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No. 10771092);国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No. 2004CB318000)。
作者简介:安娜(1983-),女,硕士研究生,主要研究领域:数据挖掘;谢福鼎(1965-),男,博士,教授,主要研究方向:人工智能、数据挖掘;张永(1975-),男,博士研究生,讲师,主要研究方向:数据仓库、数据挖掘与知识发现;刘绍海(1978-),男,硕士研究生,助教,主要研究方向:网络安全、概念网络的信息挖掘。
收稿日期:2007-11-12 **修回日期**:2008-01-28

边介数。它为区分一个社团的内部边和连接社团之间的边提供了一种有效的度量标准。按照复杂网络中社团的定义,社团内部结点之间联系紧密,而社团之间连接比较松散。所以连接社团之间的边比社区内部的边有更大的边介数。通过逐步移去这些边介数较高的边就能够把它们连接的社团分割开来。

GN 算法的基本步骤如下:

(1) 计算网络结点中所有边的介数;

(2) 找到介数最高的边并将它从网络中删除;

(3) 重复执行步骤(1),(2),直到每个节点就是一个退化的社团为止。

为了得到具有实际意义的社团结构,Newnan 等人定义了模块度[7]来衡量网络划分质量,通过计算分裂过程中每一步所产生的模块度值,最大的那个就对应了理想的社团结构。

2.2　概念格的相关理论

1982 年 R. Wille[8]首先提出了一种基于形式背景表示形式概念的新模型,即概念格。概念格又称为 Galois 格,是根据数据集中对象与属性之间的二元关系建立的一种概念层次结构,概念格通过 Hasse 图生动简洁地体现了概念之间的泛化和特化关系。

定义 1 一个形式背景(context) $K=(G,M,I)$ 由集合 G,M 及它们之间的二元关系 I 组成。集合 G 中的元素称为对象,集合 M 中的元素称为属性。对于 $\forall g\in G$, $\forall m\in M$,若 g 具有属性 m,记为 gIm 或者 $(g,m)\in I$。

假设 (G,M,I) 是一个形式背景,对于 $A\subseteq G, B\subseteq M$,定义

$A'=\{m\in M\mid \forall g\in G,(g,m)\in I\}$, $B'=\{g\in G\mid \forall m\in M,(g,m)\in I\}$

定义 2 形式背景 (G,M,I) 的一个形式概念(简称概念)是一个二元组 (A,B),它满足 $A'=B$ 且 $B'=A$,其中 $A\subseteq G, B\subseteq M$。$A$ 称为概念 (A,B) 的外延,B 称为概念 (A,B) 的内涵。

形式背景 (G,M,I) 中的概念可以用超概念和子概念的关系定义它们之间的序关系:

$(A_1,B_1)\leqslant(A_2,B_2)\Leftrightarrow A_1\subseteq A_2(\Leftrightarrow B_2\subseteq B_1)$

称 (A_1,B_1) 是 (A_2,B_2) 的子概念,(A_2,B_2) 是 (A_1,B_1) 的超概念;(G,M,I) 的所有概念的偏序集记为 $L(C,\leqslant)$,称之为 Galois 概念格,简称概念格。根据偏序关系可生成概念格的 Hasse 图,Hasse 图生动形象地再现了概念之间的关系。

定义 3 概念相似度:一个背景中的两个概念 $C_1=(A_1,B_1)$, $C_2=(A_2,B_2)$,其概念相似度定义如下:

$$Sim(C_1,C_2)=\left(\frac{|A_1\cap A_2|}{O}\right)w+\frac{\{|B_1\cap B_2|+\sum_{i=1}^{m}\max_{(1\leqslant j\leqslant n)\cap(j\notin k)}[as(b_i\in B_1\backslash B_2,b_j\in B_2\backslash B_1)]\}}{|B_1\cap B_2|+0.5\times|B_1\backslash B_2|+0.5\times|B_2\backslash B_1|\}}(1-w)\qquad(1)$$

其中,$|B_1|=m$;$|B_2|=n$;$O=\max\{|A_1|,|A_2|\}$;$k=\{j\mid as(b_s,b_j)=\max_{1\leqslant s\leqslant i-1}[as(b_i,b_j)]\}$。$w$ 是权值。一般的 $0\leqslant w\leqslant 1$ 由用户自己定义,它表明了概念中对象和属性重要程度(在本实例中,取 $w=0.2$,但在一般特殊行业,有时也可取 0,这就表明此相似度公式可分别用作对对象和属性的相似度计算)。$as()$ 表示0到1之间的一个十进制数,表示两个属性的自明相似度,这个度是在给定域中依赖专家系统统一建立的[9-11];

注:计算属性自明度之前,若 $B_1\backslash B_2$(或 $(B_2\backslash B_1)-B_1\cap B_2=\varnothing$),则两个概念之间属性自明度为属性交集的个数,无需计算其它属性自明度;

由上述公式可知,$0\leqslant Sim(C_i,C_j)\leqslant 1$,且概念的相似度具有对称性,即:$Sim(C_i,C_j)=Sim(C_j,C_i)$。

2.3　文本向量的提取

2.3.1　文本向量的空间模型(VSM)

向量空间模型是由 Salton 等人[12]在 20 世纪 60 年代提出来的,并在著名的 Smart 系统中实现。在向量空间模型中,每一篇文档被表示为规范化正交特征词矢量所组成的空间中的一个点。一般采用 IDF(Inverse Document Frequency)来表示 VSM,即:

$$d=\{(t_1,f_1\xi_1),(t_2,f_2\xi_2),(t_3,f_3\xi_3),\cdots,(t_n,f_n\xi_n)\}\qquad(2)$$

其中 t_i 为特征词项,可以是单词也可以是词组;f_i 为特征词 t_i 在 d 中出现的频率或频率函数;$\xi_i=\log(N/df_i)$,N 为本档集的文档总数,df_i 为包含特征词 t_i 的文本数,当 df_i 为 0 时,定义 $\log(N/df_i)$ 为 0。

2.3.2　文本权值的计算

使用 TFIDF 方法来计算特征词的权重时,不能准确地反映词汇在文章中的重要程度[13],因为特征词的权重是由许多因素决定的,例如特征词的词长、特征词的位置、受限语义的分析等,所以,给出特征词权值的计算公式;设特征词 t_i 的权值为 W_i,则有:

$$W_i=\delta\times\mu\times f_i\times\log(l+1)\qquad(3)$$

其中,δ 为特征词位置的加权系数;μ 为受限语义分析加权;f_i 为特征词在文档集中出现的频率;l 为特征词的长度。

2.3.3　文本特征向量的调整

通过以上步骤得到的文档特征向量是一个超高维稀疏向量,这是很不利于聚类的,因此还需降低特征空间的维数,首要的办法是从特征空间中选取最具代表性的特征词作为文档的特征空间向量。

步骤如下:

(1)删除在文档集中词频较低的词。对于一个特征词 t_i,用 dt_i/N 表示该词在文档集中的出现率。dt_i 表示包含该词的文档数;N 表示文档集总数。当 dt_i/N 小于一定值时,则删除该特征词项。

(2) 删除文档集中分布均匀的词。方差 $\sigma_i=\sum(m(k,i)-\frac{1}{N}\sum_{j=1}^{N}m(j,i))^2$ 表示词 t_i 在文档集中的分布情况。$\boldsymbol{M}$ 表示文档集的特征矩阵,N 表示文档集总数,σ_i 越小,表明该词分布均匀,所以 σ_i 小于一定值时,则删除该特征词项。

综合以上分析,特征词 t_i 的衡量值为 $\eta_i=\frac{dt_i\times\sigma_i}{N}$,当 η_i 小于一定值时,表示该词不利于聚类,应该删除。

3　基于 GN 算法的文本聚类方法

本文给出的文本聚类方法是概念格和复杂网络两种理论相结合的聚类方法,首先对待聚类的文档集进行预处理,提取并计算各文档关键词的特征权值;其次构造形式背景,形式背

景的行为文档标号,列为文档的所有关键词,它反映了关键词与文档的匹配结果;应用建格算法构造所有概念,从中选取具有代表性的概念,选取特征概念的标准为所有特征概念的对象之和为对象全集,属性之和为属性全集,为了达到分类精确,应尽量选取对象较少的概念,所以本文应用一种概念选取算法来得到最终的特征概念;最后,通过本文所给出的相似度公式,计算出形式背景中概念相似度的大小,应用 GN 算法进行文本概念聚类。

基于 GN 算法的文本聚类算法描述如下:

输入:待聚类的文档;阈值 λ(限定关键词权值的变量);

输出: 文本聚类结果;

步骤 1 计算并提取每篇文档各关键词的特征向量权值;从专家库中读取关键词的自明度 $as()$;

步骤 2 构造形式背景 (G,M,I),将权值大于 λ 的关键词在所对应的文档处标 1,否则不作任何标记;

步骤 3 构造概念。依据形式背景应用建格算法构造所有概念。选定具有代表性的部分概念 C_i 进行计算,使得 $\cup C_{i.}\,g = G \cup C_{i.}\,m = M$,其中,$C_i$ 表示第 i 个概念;$C_{i.}\,g$ 表示第 i 个概念的对象,$C_{i.}\,m$ 表示第 i 个概念的属性;

应用如下算法抽取特征概念:

首先将所有概念按对象数目由少到多进行排列;

设概念 $C(g,m)$,$C(g,m)$ 表示算法执行过程中所抽取的概念的所有对象和所有属性之和的概念;

```
设概念 C(g,m) C. g = ∅; C. m = ∅;
for(j = 1; j≤n; j++ ) { n 表示所有概念数}
  { if( C. g + C_j. g < G && C. m + C_j.m < M )
    {把 C_j(g,m) 存入到所选概念的数组中;
    C. g = C. g + C_j. g;   C. m = C. m + C_j. m;
      j++;
    }
  else if( C. g + C_j. g == G || C. m + C_j. m == M)
      j++;
    else( C. g + C_j. g == G && C. m + C_j. m == M)
        exit;
  }
```

步骤 4 根据公式(1)计算各概念间的相似度并构造相似矩阵 $\boldsymbol{R} = (Sim(C_i,C_j))_{n\times n}$;

步骤 5 构造 GN 矩阵 $\boldsymbol{L}$。选取个阈值 θ 对矩阵进行 θ-截集处理,相似度大于 θ 的转化为 1,小于 θ 的转化为 0,形成新的矩阵 $\boldsymbol{L}$,(容易证明,θ 越大分出的类别数越多,θ 越小分出的类别数越少,特别当θ=0时分得最粗,当 θ=1 时分得最细。利用 θ 取值不同可得出不同程度的分类结果):这样得出 GN 矩阵 $\boldsymbol{L} = (Sim(C_i,C_j))_{n\times n}$;

步骤 6 应用 GN 算法得出概念集的聚类结果。

4　例子

在这一部分中,通过实例来说明所给算法的计算过程。形式背景如图 1。

应用建格软件,将上述形式背景转化成 Hasse 图,Hasse 图共有 5 层,95 个概念。本文选出 10 个代表性的概念,具体表述如下:

x_1(2 5 7 13 ,ack) x_2(14 16,adj) x_3(15 16 20 21,$afgj$) x_4(4 8 9 ,$abei$) x_5(1 9 11,$abio$) x_6(1 3 11,$adio$) x_7(11 16,$adfgo$) x_8(17 19,$agjln$) x_9(2 9 ,$acfhi$) x_{10}(10 11 12,$abdgn$)

自明度为:$as(eg)=0.9$;$as(cd)=0.9$;$as(ki)=0.9$;$as(dl)=0.9$;$as(ij)=0.9$;$as(fn)=0.9$;$as(eo)=0.8$;$as(kj)=0.8$;$as(fk)=0.8$;$as(co)=0.7$;

由于本例相似度的计算主要应用属性的匹配,所以将属性的权值设为 0.8,则对象权值为 0.2。

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
1	1	1		1					1				1		1
2	1		1			1		1	1	1	1				
3	1			1					1						1
4	1	1					1		1			1			
5	1	1	1			1			1		1				
6	1	1		1					1				1	1	
7	1	1	1				1		1		1				
8	1	1	1						1						
9	1	1	1			1	1	1	1						1
10	1	1		1			1	1	1					1	
11	1	1		1		1	1	1	1					1	1
12	1	1		1			1		1					1	
13	1		1							1	1				
14	1			1		1				1		1			
15	1					1	1			1					
16	1			1			1			1			1		1
17	1						1	1		1		1		1	
18	1						1			1				1	
19	1						1			1		1		1	
20	1	1				1	1	1		1			1		
21	1	1				1	1	1		1			1		

图 1　21 个文档的形式背景

基于上述准备工作,应用相似度公式(1)得到图 2 特征概念相似度矩阵 $\boldsymbol{R} = (Sim(C_i,C_j))_{10\times 10}$:

$$\boldsymbol{R}=\begin{Bmatrix}
1 & 0.72 & 0.41 & 0.43 & 0.59 & 0.64 & 0.54 & 0.36 & 0.63 & 0.38\\
 & 1 & 0.51 & 0.43 & 0.43 & 0.66 & 0.50 & 0.52 & 0.56 & 0.40\\
 & & 1 & 0.56 & 0.38 & 0.38 & 0.58 & 0.69 & 0.52 & 0.50\\
 & & & 1 & 0.76 & 0.56 & 0.34 & 0.50 & 0.42 & 0.52\\
 & & & & 1 & 0.73 & 0.42 & 0.43 & 0.55 & 0.42\\
 & & & & & 1 & 0.60 & 0.50 & 0.48 & 0.42\\
 & & & & & & 1 & 0.61 & 0.46 & 0.55\\
 & & & & & & & 1 & 0.30 & 0.57\\
 & & & & & & & & 1 & 0.45\\
 & & & & & & & & & 1
\end{Bmatrix}$$

图 2　特征概念相似度矩阵

本例设 θ=0.52,对矩阵进行 θ=0.52 截处理。应用 GN 算法将 10 个概念聚类。图 3 表示为以 x_8 为顶点的边介数的计算,按结点所在的层数从左到右的顺序,结点依次为($x_8x_3x_7x_{10}x_4x_1x_6x_9x_5x_2$)。将 10 个概念分别作为源点计算边介数,最后将每次所得的权值相加,就可以得到所有边的总介

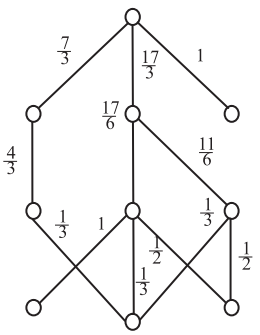

图 3　x_8 为源点的边介数计算

(下转 180 页)