《智能信息处理》课程作业

基于形式概念分析的文本聚类研究

孔玉

作业	分数[20]
得分	

2020年11月11日

基于形式概念分析的文本聚类研究

孔玉

(大连海事大学 信息科学技术学院 辽宁 大连 116026)

摘 要： 形式概念分析是 1982 年由德国的 Wille 提出的一种从形式背景进行数据分析和规则提取的强有力工具，已被广泛地研究，并应用到软件工程、知识发现、信息检索等领域。文本聚类是聚类方法与文本处理中的自然语言处理相结合，是自然语言处理领域的一大研究热点。鉴于文本聚类面临的传统向量空间模型特征维数过高、稀疏向量以及标准 K-Means 算法初始中心点选择的随机性等问题，本文提出将形式概念分析应用于文本聚类中，利用概念来表示文档，最后考虑基于改进的 K-Means 算法进行文本聚类的方式，使特征词的维数降低，减少计算的复杂度并且提高计算的精确度。

关键词： 形式概念分析;自然语言处理;文本聚类;K-Means 算法

中图法分类号 TP391 **文献标识码** A

Text Clustering Based on Formal Concept Analysis

Kongyu

(Institute of Computer Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract: Formal concept analysis is set up in 1982 by the German Wille proposed a form from the background of the powerful tool for data analysis and rule extraction, has been widely studied, and applied to the software engineering, knowledge discovery, information retrieval, and other fields. Text clustering is a research hotspot in the field of natural language processing, which combines the clustering method with the natural language processing in text processing. Consideration of text clustering traditional vector space model is too high, the characteristic dimension of the sparse vector and the standard K - Means algorithm initial center selection problem such as randomness, in this paper, the formal concept analysis is applied to text clustering, use concepts to represent document, based on the modified K - Means algorithm for text clustering methods, reduce the dimension of key, reduce the complexity of the calculation and improve the accuracy of the calculation.

Key words: Formal concept analysis; Natural language processing; Text clustering; K - Means algorithm

0 引言

聚类(Clustering)又称聚类分析(Clustering Analysis) [1,2]，是最重要的无监督学习的方法。聚类是一个将数据集划分为若干类的过程，并使得同一个类的数据对象具有较高的相似性，而不相同类中的数据对象则具有较大的相异性。

文本挖掘属于数据挖掘这一交叉学科的一个具体领域，它的主要任务是分析文档数据库的内容，发现文档数据集中概念、文档之间的相互关系和作用，抽取有效、新颖、有用、可理解的、 散布在文本文件中的有价值的知识，并利用这些知识更好地组织信息。文本挖掘处理的是非结构化的文本信息，而不是通常数据挖掘中采用的结构化数据信息。文本挖掘的主要研究内容包括关 联分析、文本分类、文本聚类等。

在对海量中文文本进行聚类等分析时，常用向量空间模型(VSM)来表示文本，对文本进行过滤停用词、去噪等操作后，特征维数依然很大[3]。与传统的聚类方法不同的是，基于概念的聚类是利用概念集来进行聚类。目前中文概念集的生成主要有两种方法：基于概率分布的概念生成算法[4] 和基于自组织映射神经网络的概念生成算法(Self-Organizing Map, SOM)[5]。然后再利用概念集簇来进行聚类、检索等。但是这两种算法都有 复杂度较大，难以适应大规模的文本聚类的缺点。

形式概念分析是一种从形式背景进行数据分析和规则提取的强有力工具，它建立在数学基础之上，概念格是其核心数据结构。作为数据分析和知识处理的形式化工具，概念格理论已被广泛地 应用于软件工程、知识发现、数据挖掘和信息检索等领域。本文就是利用概念格理论结合改进的K-Means算法来实现文本聚类的。

1 理论介绍

1.1 研究意义及背景

21世纪，计算机技术和网络通信技术正在推动人类各方面的进步，互联网已经成为人们不可缺少的信息来源。目前，网络资源数据增长速度飞快，人们要获取所需的信息要花费很多时间， 所以，如何快速准确获取信息成为焦点。传统的信息搜索技术存在着这局限性，已经不能适应目前增加的大量文本数据处理，文本挖掘(Text Mining)成了数据挖掘的一个很有前途的研究方向。文本处理的特殊性，不能像数据库中的数据， 文本处理需要有自然语言理解的支持，目前机器对自然语言理解还存在很多歧义问题，因此文本挖掘还不能很好的表达理解的层次。[6]

概念格是一组概念的序集，建立概念格的过程就是对概念进行聚类的过程。在概念格中，概念的外延为属于这个概念的所有对象的集合，而内涵是所有这些对象所共有的属性集。给定一个形式背景就能在此基础上构造概念格，且构造出的概念格是唯一的。作为数据分析和知识处理的形式化工具，概念格理论已被广泛地应用于软件工程、知识发现、数据挖掘和信息检索等领域。

1.2 形式概念分析

形式概念分析(Formal Concept Analysis, FCA)是由德国的 Wille 教授于 20 世纪 80 年代初提出的[8]，它反映了概念的哲学理解，其核心数据结构概念格，也称 Galois 格，准确而简洁地描述了概念之间的层次关系，因此成为一种重要的知识表示方法。随着研究的深入，形式概念分析越来越多地被应用到数据挖掘、信息检索、软件工程等领域，成为处理和组织大规模数据的有效工具。本节主要介绍概念格模型的一些基本概念和相关的研究内容，并总结了形式概念分析的应用现状。

1.2.1 概念格

概念格[7](concept lattice)是 FCA 的核心数据结构。概念格的每个节点是一个概念，每个概 念由外延（Extent）和内涵(Intent)两部分组成。外延是概念所覆盖的实例，而内涵是概念的描述， 是该概念所覆盖实例的共同特征。概念格可以通 过其 Hasse 图生动简洁地体现概念之间的泛化和例化关系。因此，概念格被认为是一种支持数据 分析的有效工具。

概念格的构造过程实际上是概念聚类的过程，是我们使用形式概念分析进行处理问题的前提。通常，概念格的大小是在指数量级上的，而且要处理的数据又多数是海量的，概念格构造算法的研究始终是形式概念分析中的一个主要问题。概念格的构造算法可分为两类:批处理算法和渐进式算法。[8]

在最坏情况下，生成的概念格呈现指数级爆炸式增长。因此在数据量很大的情况下，控制格 中节点的增长是必须的。这时需要使用一定的方法进行约简。常用的属性约简[9]的方法有可决策表属性约简，可辨识矩阵法等。决策表属性约简 中，一个决策表就是一个决策信息系统，其中包 含了大量领域样本实例的信息。决策表中的一个 样本就代表一条基本决策规则，所有的决策规则 的集合就形成一个决策集，用于属性约简。可辨 识矩阵法是由斯科龙教授[10]提出的，应用可辨识矩阵能够得到全部的可约简组合，也就是说能够获得理论上的最优约简。

1.2.2 形式背景

一个形式背是一个三元组 $FC=(G,M,I)$，其中G、M 是非空有限集合，G是对象的集合，M 是属性的集合，I是G 和M 之间的二元关系，对于$\forall g \in G$，$m \in M$，若$(g,m) \in I$，就说g 具有属性m，记做 $g\,\mathrm{Im}$ 。

一个形式背景实际上与数据库中的一张二维表类似（如表 1 所示），对象相当于表的记录，而属性则相当于表的字段，对象与属性之间的关系相当于表中记录与字段的关系。

1.2.3 形式概念分析的应用

作为数据分析和知识处理的形式化工具，形式概念分析已经被应用于越来越多的现实研究中。在形式概念分析的研究中，所有数据是以概念格的形式进行有机的组织，而概念格则体现了概念内涵与外延的统一，因此对一些规则类型的知识获取很有帮助。概念格现已被广泛的应用于

自然语言处理、知识库组织等诸多领域。

1.3 文本聚类

聚类源于很多领域，包括数学，计算机科学， 统计学，生物学和经济学。K-Means算法是一种传统的聚类算法，是Mac Queen在1967年提出的到目前为止在聚类算法中比较有影响力的算法 技术。该算法具有聚类速度快、易于实现等特点， 是典型的基于距离的聚类算法，采用距离作为i 相似性的评价指标，将n个对象按K个簇划分，要求簇与簇之间相似度低，即距离越远越好，而每个簇内部相似度极高，最终把得到紧凑而独立的簇为实现目标。因此文本聚类在处理问题方面具
有很强的灵活性。以下为几种常用的聚类算法[11]：
(1) 划分方法（partitioning methods）
(2)层次方法（hierarchical methods）
(3)基于密度的方法 (density-based methods)
(4)基于网格的方法 (grid-based methods)
(5)基于模型的方法 (model-based methods)

2 基于形式概念分析的文本聚类

2.1 文本预处理与特征词选取

对文本进行预处理操作，包括去掉标签(Tag)、停用词过滤、词缀剪枝(Stemming)以及归类组织等。其中停用词过滤只需要用 Stopwords 算法[13]的停用词列表(Stoplist)就可以完成，而词缀剪枝 可以采用 Martin Potter 博士的 Porter 词干分析算法[14]，此算法是对英文中因时态、 语态、复数等原因引起的词尾变化进行移除的处理过程，可
将每个单词的各种形式还原为词干原型。

特征词的选取是文本聚类的基础，常见的特征词选取方法有以下几种：根据词频抽取特征，如 LIRA 和 DICA； 根据词频/ 倒置文档频率(TF-IDF)提取特征，如 TF*IDF；根据互信息量提取特征； 根据期望信息增量提取特征， 如 Syskill&Webert 的研究； χ^2统计(CHI)方法。

2.2 用概念表示文本，生成形式背景

现在文本的表示大多是基于向量空间模型（VSM）[15]的。文档被表示 t-维空间向量。对 文本的表述文档集 $D=\{d_1,d_2,d_3,...,d_n\}$。其中 $d_1=<(t_1,w_1),(t_2,w_2),...,(t_n,w_n)>$。$\boldsymbol{w}_i$为此特征词$t_i$在文档$d_i$中的权重，权重的大小可以由公式 TF*IDF 来计算。

即 ：

$$tfidf\ (t_i,d_i)=tf(t_i,d_i)*Log\ \frac{|T_r|}{|T_r(t_i)|}$$

$$tf(t_i,d_i)=1+Log(N(t_1,d_i))$$

其中T_r表示文档总数，$T_r(t_i)$表示在T_r中出现的t_i的文档数，$N(t_i,d_i)$ 表示特征词t_i在文档d_i中出现的次数。

这种方法是基于特征词出现的频数tf，没有充分考虑文档特征词之间可能存在的语义关系，两个文档之间虽存在语义关系，但由于出现的次数不同，而被归于不同的类。更糟糕的是用 VSM 表示文档会导致文本特征词的维数过高，从而使计算的复杂度增加。但利用概念来表示文档能很好的解决这个问题。因为概念中包含多个属性，与 VSM 相比维数有所降低，从而降低了计算的复杂度。设定文档为概念格中的对象集，文档中的关键词或术语(Term) 构成属性集。从而得到基于文档的形式背景的定义：一个基于文档的形式背景是一个三元组 $K=(D,T,I)$，其中 D是文档（对象）集，T是关键词（属性）集， I是 D和T之间的二元关系，它表明文档 d 中是否有关键词 t。如果 t 是 d 的关键词， 则记为 dIt 或 $(d,t)\in I$。

如表 1 表示一个形式背景，文档集 D={ d1, d2, d3, d4, d5}，关键词集 T={ a, b, c, d} ，表 1 中的“1”表示文档中有该关键词，“0” 表示文档中没有该关键词。

表1 形式背景示例

	a	b	c	d
d1	1	1	1	1
d2	1	1	0	0
d3	0	1	1	1
d4	0	1	0	0
d5	0	1	1	0

2.3 概念格构造及属性约简

2.3.1 概念格的构造

下面以表 1 为例，介绍一下如何通过形式背景构造概念格。

定义在形式背景 $K=(D,T,I)$中，若对象集 $A\in P(D)$，属性集$B\in P(T)$之间按如下关系连接 ： $f(A)=\{t\in T\mid d\in A,tId\}$ ，$g(B)=\{d\in D\mid t\in B,tId\}$。则称从形式背景中得到的每一个满足 $A=g(B)$，$B=f(A)$的二元

组(A,B)为一个形式概念。其中 A 是对象密集 $P(D)$的元素，称为概念(A,B)的外延， B 是属性密集 $P(T)$的元素，称为概念(A,B)的内涵。

利用经典的构格算法[16,17]构造表 1 所示形式背景的概念格，相应的 Hasse 图如下：

C1 ({d1, d2, d3, d4, d5}, {b})

C2 ({d1, d3, d5}, {a, b}) C3({d1, d3}, {b, c, d})

C4 ({d1, d2}, {a, b}) C5 ({d1}, {a, b, c, d})

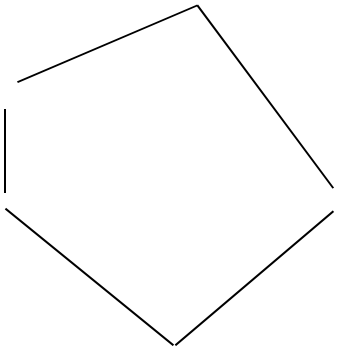

图 1 基于表 1 形式背景的哈斯图

表 2 对图 1 节点做相应解释

概念	内涵	外延
C1	b	d1,d2,d3,d4
C2	a,b	d1,d3,d5
C3	b,c,d	d1,d3
C4	a,b	d1,d2
C5	a,b,c,d	d1

如图 1 所示，该概念格包含 5 个概念，概念用 C 来表示，分别为：C1 = ({ d1, d2, d3, d4, d5} , { b}) ，C2 = ({ d1, d3, d5} , { b, c})，C3= ({ d1, d3} , { b, c, d})，C4= ({ d1, d2} , { a, b})，C5= ({ d1} , { a, b, c, d})。所有概念的集合 C={C1, C2, C3, C4, C5}。对于任何两个概念之间的相似程度，通过定义的相似度函数计算。设 Ci= (A, B)，其中A是概念Ci的外延，B是概Ci的内涵。Cj=C，表示其中的一个概念。概念之间的相似度函数可以由以下定义确定：

定义

$$sim(C_i, C_j) = (1-w) * \left|\frac{A \cap Extent(C)}{A \cup Extent(C)}\right| + w * \left|\frac{B \cap Intent(C)}{B \cup Intent(C)}\right|$$

其中 w 为赋予的权值，范围在 0 与 1 之间。可以通过调整 w 值的大小来反映属性的重要程度。当两个概念 Ci 和 Cj 没有任何相似对象和属性时，sim(Ci,Cj) =0；当Ci 和Cj 完全相同时 sim(Ci, Cj) =1，sim(Ci, Cj)的数值越接近 1 时，两个概念之间的相似度越大；反之，数值越接近 0 时，两个概念的相似度越小。

从概念中很容易得出和文档相关的概念，若概念中包含该文档就说概念和该文档是相关的。如概念 C3 的对象有 d1 和 d3 两篇文档，即文档 d1 和 d3 与概念 C3 都是相关的，从而可以得到与每个文档相关的所有概念，如表 3 所示。

表 3 文档及关联概念

d1	C1	C2	C3	C4	C5
d2	C1	C4			
d3	C1	C2	C3		
d4	C1				
d5	C1	C2			

表 3 列出了表示文档的每个概念，利用概念表示文本，通过概念的相似度可以度量文本之间的相似度。文本相似度函数由以下定义确定：

定义[9]

$$sim(d_i, d_j) = \left|\frac{1}{k}\right| \sum_{i=1}^{m} \sum_{j=1}^{n} sim(C, C)'_j$$

其中 m 为文档 di 中包含的概念个数，n 为文档 dj 中包含的概念个数，k=m*n 为两个文档中概念个数的乘积。例如：在概念相似度函数定义中 取 w=0.5， 则 sim(d4, d5) = 12(sim(C1, C1) +sim(C1, C2)) =12(1+0.55) =0.7750，sim(d2 , d5) = 14(sim(C1 , C1) +sim(C1 , C4) +sim(C2 , C1) +sim(C2, C4)) =0.5829。

2.3.2 概念格的属性约简

随着聚类文本对象和属性的增加，形式背景以及格结构都变的非常庞大和复杂，这对于聚类 结果的准确度危害很大，为了提高聚类的准确度以及降低聚类的复杂度，在聚类之前就要先对概念格进行属性约简，降低形式背景的维度，使形式背景更加简洁。本文所使用的属性约简算法是 基于可辨识矩阵法，在此方法中，只有非空元素对此约简有意义。经过约简后，最终形成的矩阵是一个对称矩阵，且该矩阵在对角线上的元素均为Ø。

基于可辨识矩阵的属性约简[7]的步骤如下：

Step1：计算决策表的可辨识矩阵，并将矩阵中属性组合数为的属性列入最终的属性约简集合。

Step2：对于所有取值为非空集合的元素，建立相应的析取逻辑式

Step3：将所有的析取逻辑表达式进行合取de得到一个合取范式，并将合取范式转化为析取范式的形

式。

Step4：输出属性约简结果，析取范式中的每个合取项就对应一个属性约简的结果，每个合取项中所包含的属性组成约简后的条件属性集合。

定义[9] 对于一个形式背景 $K=(D,T,I)$，$D=\{d_1,d_2,...,d_n\}$，$T_i(d_j)$表示对象d_j在属性T_i的取值(0 或 1)。$C_D(i,j)$表示可辨识矩阵第i行 j 列的元素：

$$C_D(i,j)=\{t_k | t_k\in M\wedge t_k(d_i)!=t_k(d_j)\}$$

其中$i,j=1,2,...,n$ 。

则由该定义可求出表 1 的可辨识矩阵为：

$$\begin{bmatrix} 0 & cd & a & \text{acd} & ad \\ & 0 & acd & a & ac \\ & & 0 & \text{cd} & d \\ & & & 0 & c \\ & & & & 0 \end{bmatrix}$$

进而约简后的形式背景为：

表 4 约简后的形式背景

	a	c	d
d1	1	1	1
d2	1	0	0
d3	0	1	1
d5	0	1	0

3 评价与改进

本文主要是从形式背景中得到的所有概念，采用枚举的方法选取概念来表示文本。如何选择出最合适的概念来表示文本，提高文本表示的准确率，将是以后研究的一个重点。由于在文本形式背景中文档和特征词的关系是用 1 或0 来表示，有其局限性，用模糊形式背景来表示文档，文档和特征词的关系通过定义的隶属度函数来度量将是一个改进方向。但未能实现用 K-MEANS 算法进行改进，以后可以逐步完善。

参考文献

[1] Jiawei H, Kamber M. Data mining: concepts and techniques [J]. San Francisco, CA, itd: Morgan Kaufmann, 2001, 5.
[2]数据挖掘: 概念与技术[M]. 机械工业出版社, 2007.
[3]李江华 杨书新 刘利峰. 基于概念格的文本聚类[J]. 计算机应用, 2008, 28(9): 2328-2330.
[4]宫秀军, 史忠植. 基于 Bayes 潜在语义模型的半监督 Web 挖掘[J]. 软件学报, 2002, 13(8):1508-1514.
[5]朱晓敏. 基于概念格扩展模型的无标签文本挖掘方法研究[D].2018.
[6]唐明珠.形式概念分析和本体在文本挖掘中的 应用[D].兰州理工大学,2008.
[7]吴湘华, 曹丽君. 可变属性粒度的中文文本概念格聚类研究[J]. 电脑知识与技术:学术版, 2019(26)
[8]R.Wille.Restructuring lattice theory:an approach based on hierarchies of concepts[J].in Ordered Sets , I.Rival , Ed.Reidel, Dordrecht, 1982, pp: 445-470.
[9]李云, 田素方, 李拓, 等. 基于概念格的 Web 文本聚类[J].计算机工程与应用,2009,44(23)
[10]刘萍, 彭小芳. 基于形式概念分析的词汇相似度计算[J]. 数据分析与知识发现, 2020, v.4;No.41(05):70-78
[11] Skowron A. Extracting laws from decision tables: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 371-388.
[12] Kim M, Compton P. Evolutionary document management and retrieval for specialized domains on the web [J]. International Journal of Human-Computer Studies, 2004, 60(2): 201-241.
[13] Porter M F. An algorithm for suffix stripping [J]. Program, 1980, 14(3): 130-137
[14] Berry M W, Drmac Z, Jessup E R. Matrices, vector spaces, and information retrieval[J]. SIAM review, 1999, 41(2): 335-362.
[15] Godin R, Mineau G, Missaoui R. Incremental structuring of knowledge bases[C]//Proc. of KRUSE. 1995, 95: 179-193.
[16] 和晓萍 HXP. 一种优化的 k -Means 聚类中心初始化方法[J]. 云南民族大学学报(自然科学版), 2015, 24(S).
[17]李艳霞, 史一民, 李冠宇. 基于概念格的 K—Means 算法研究[J]. 计算机工程与设计, 2011, 32(2): 656-658.
[18]韩永花. 基于概念格的多文本知识源挖掘[D]. 2016.