22届机检会论文选登

文本聚类结果描述研究综述*

章成志
（中国科学技术信息研究所　北京 100038）
（南京理工大学信息管理系　南京 210094）

【摘要】首先对文本聚类结果描述的研究背景和相关的研究情况进行说明，分析自动标引、自动文摘、概念聚类与文本聚类结果描述的关系，定位文本聚类结果描述的研究内容；然后根据文本聚类结果描述的具体要求，对该问题进行形式化；最后给出文本聚类结果描述的评价方法。

【关键词】文档聚类描述　文本聚类　文本挖掘
【分类号】TP391　G252

Survey on Document Clustering Description

Zhang Chengzhi
(Institute of Scientific and Technical Information of China, Beijing 100038, China)
(Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094, China)

【Abstract】The research background and related research work about Document Clustering Description (DCD) are given in this paper. The relationship between DCD and automatic indexing, automatic summarization, conceptual clustering is explained and the research content of DCD is definited. According to its requirements, the tasks of DCD are formalized. The evaluation methods of DCD are also described in this paper.

【Keywords】Document clustering description　Document clustering　Document mining

1 引　言

标注文档集合聚类后生成的类簇，可以让用户更容易通过类簇的标签来了解各个类簇的主题，节省信息浏览时间。通常，该任务被称为文本聚类结果的类簇标注（Cluster Labeling）[1-4]、类别标注[4]、类簇命名（Cluster Naming）[5,6]、标签识别（Label Identification）[4,7]、主题发现或识别[4]、文本聚类结果描述（简称聚类描述，Cluster Description，如无特殊说明，本文后面提到的聚类描述特指文本结果聚类描述）[8]、描述聚类（Descriptive Clustering）[8]或文本聚类结果类别标题的自动生成（Title Generation for Clustered Documents）[9]。在机器学习和数据挖掘领域，聚类描述是概念聚类的后处理部分[10]。

现有的文本聚类方法中共存的问题是聚类结果的有效描述问题，也是查询结果聚类（Search Results Clustering）的难点问题之一。传统的聚类算法直接用于文本聚类上，存在的突出问题就是算法的有效性问题，因为传统的聚类算法只对对象进行聚类，不负责对聚类后生成的类簇进行概念描述和解释。因此，必须针对文本聚类的特

收稿日期：2008－11－18
* 本文系中国博士后科学基金资助项目“多语领域本体学习关键技术研究”（项目编号：20080430463）、南京理工大学科研启动基金项目“主题聚类关键技术研究”（项目编号：AB41123）和“十一五”国家科技支撑计划重点项目“多语言信息服务环境关键技术研究”（项目编号：2006BAH03B02）的研究成果之一。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

别要求，探寻专门解决文本聚类描述这一问题的方法。

聚类描述是帮助用户迅速确认生成的文档类相关与否的重要信息。聚类描述是一项很具有挑战性的工作[7]，具有重要的研究意义和应用价值。本文分析自动标引、自动文摘、概念聚类与文本聚类结果描述的关系，定位文本聚类结果描述的研究内容；然后根据文本聚类结果描述的具体要求，对该问题进行形式化。最后给出文本聚类结果描述的评价方法。

2 相关研究工作和研究领域分析

2.1 相关研究工作

按照聚类描述生成的自动化程度，可将其分为人工描述方法和自动描述方法。Patrick G. & Wolfgang G., Lai & Wu等人通过人工描述方法完成聚类描述工作[11,12]。自动化的聚类描述，主要从聚类生成的类簇中提取重要的词语，根据聚类算法的不同，相应的词语重要性计算方法也有所不同[9]。下面对几种有代表性的聚类描述方法做介绍。

1992年，Cutting, Karger & Pedersen等人在Scatter/Gather系统中，利用归一化的词语频率作为词语权重，选择权重较大的词语组成列表作为类别描述[13,14]。后来，Muller A. & Dorre J.等人则直接将聚类类簇的前N个最高频次的词语作为聚类描述[15]。此类方法存在的弱点在于，选择单一的词语或者列表作为类簇描述，存在可读性和可理解性的问题。1996年，Anton V. Leouski & Croft W. B.等人则用关键短语来进行类别描述，认为基于短语的聚类描述优于基于单个词的聚类描述[16]。1998年，Zamir & Etzioni将类簇的文档集中的出现频次(TF)高的最长短语作为聚类描述[17]。2001年，Lawrie D. & Croft W. B.等人从聚类生成类簇的成员中提取TF×IDF值大的词语作为聚类描述[5]。值得注意的是，这种简单的基于统计的方法难以从根本上提高类簇描述的查准率。2002年，Glover E. & Pennock D. M.等人通过对文档集合中的父子、同类特征建立统计模型，通过层次关系进行推理得到聚类描述[18]。需要指出的是，在层次类簇中，无论是父类、类自身还是子类，给出的类簇为词语列表形式，可读性不强。

2006年，Tseng & Lin等人借助WordNet作为外部资源提取类别词作为聚类描述符，并利用上位词搜索算法将类别描述符转换为宽泛的、可作为聚类描述的词语[9]。如图1所示，他们提供给用户的聚类描述是重要词语列表形式，即最初提供由5个单词构成的列表，经过WordNet转换后，最后提供由三组单词并附带每组权重的列表，依然存在可读性不强的问题。另外，利用WordNet作为外部资源提取聚类的类簇描述词，存在未登录词问题。但提出的利用相关系数获取候选类簇描述词的方法是值得借鉴的。同年，Puck tada Treeratpituk & Jamie Callan综合利用描述词在类簇本身、父类簇和文本集合的统计信息来对描述词进行描述能力打分，如描述词的文档频率DF，TF×IDF及它们的排序信息等，最终得到每个类簇的标注[2]。需要指出的是，若针对篇平类簇结构(即类簇无层次结构)的类簇进行描述，则无父类簇信息可以参考。另外，Puck tada Treeratpituk & Jamie Callan给出的聚类描述结果依然是多个单词组合的形式，依旧存在可读性问题。亦于同年，Dawid Weiss提出了DCF(Description Comes First)算法来解决传统方法中存在的聚类描述可读性不强的问题，与传统的聚类结果描述算法不同的是，DCF算法是在文本聚类完成的同时，聚类标签也生成出来。该方法生成的聚类标签是类簇中心向量的替代物[8]。由于Dawid Weiss是采用事先生成标签来代替聚类后生成的类簇中心向量，这样一方面使得聚类描述与聚类类簇的中心向量之间存在一定的“语义间隔”，另一方面，这与“先聚类、后描述”的直觉相违背，减低了聚类描述的可解释性。本文拟综合利用“先描述、后聚类”和“先聚类、后描述”的优点，解决可理解性问题。

ID	Cluster's Descriptors	WordNet	InfoMap
1	acid, polymer, catalyst, ether, formula	1:substance, matter:0.1853 2:drug:0.0980 **3:chemical compound:0.098**	**1:chemical compound:1.25** 2:substance, matter:1.062 3:object, physical object:0.484
2	silicon, layer, transistor, gate, substrate	1:object, physical object:0.1244 2:device:0.1211 3:artifact, artefact:0.1112	1:object, physical object:0.528 2:substance, matter:0.500 3:region, part:0.361
3	plastic, mechanism, plate, rotate, force	1:device:0.1514 2:base, bag:0.1155 3:cut of beef:0.1155	1:device:0.361 2:entity, something:0.236 3:chemical process:0.0
4	output, signal, circuit, input, frequency	**1:communication:0.1470** **2:signal, signaling, sign:0.1211** 3:relation:0.0995	**1:signal, signaling, sign:1.250** **2:communication:1.000** 3:abstraction:0.268
5	powder, nickel, electrolyte, steel, composite	1:substance, matter:0.1483 **2:metallic element, metal:0.1211** 3:instrumentation:0.0980	**1:metallic element, metal:0.500** 2:substance, matter:0.333 3:entity, something:0.203
6	gene, protein, cell, acid, expression	1:substance, matter:0.1112 2:object, physical object:0.0995 3:chemical compound:0.0980	1:entity, something:0.893 2:chemical compound:0.500 3:object, physical object:0.026

图1 一个典型的聚类描述结果示例[9]

2.2 相关的研究领域

如图2所示，与聚类描述相关的研究领域包括文本聚类、自动标引、自动摘要、概念聚类、话题检测以及本体构建等。其中，自动标引、自动摘要、概念聚类与

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

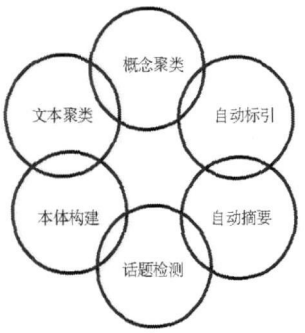

图 2　与聚类描述相关的研究领域

文本聚类结果描述的关系最为密切，本文着重对他们之间的关系进行阐述。

（1）文本聚类结果描述与自动标引

自动标引是从文本中提取出能反映文本的主题的关键词或者主题词。根据标引的词语的来源不同，可以将自动标引分为抽词标引和赋词标引。

文本聚类结果描述可以看作多文档的自动标引，即利用关键词或主题词从主题来描述多个文本的主要内容。根据多个文档涉及到的主题分布情况，可以将多文档自动标引分为多文档单主题自动标引和多文档多主题自动标引两种标引情形。文本结果聚类描述与自动标引的关系如表 1所示：

表 1　文本结果聚类描述与自动标引的关系

	单主题		多主题	
	抽词	赋词	抽词	赋词
单文档	单文档单主题抽词标引	单文档单主题赋词标引	单文档多主题自动抽词标引，即传统的抽词标引	单文档多主题自动赋词标引，即传统的赋词标引
多文档	多文档单主题抽词标引	多文档单主题赋词标引	多文档多主题自动抽词标引	多文档多主题自动赋词标引
	传统文本聚类结果描述	单个描述词经外部资源转换的聚类描述	一个类簇有多个描述词的文本聚类结果描述	多个描述词经外部资源转换的聚类描述

本文的文本聚类结果描述针对的是聚类后生成的每个类簇的主题描述，通常是单个主题的描述，因此可以将文本聚类结果描述进一步简化为多文档单主题自动标引。

不同于一般自动标引的是，文本聚类结果描述是解决多文档单主题自动标引问题所使用的训练集，所使用的特征与一般自动标引中使用的特征有所不同。根据聚类描述用词的来源不同，可以将多文档单主题自动标引分为多文档单主题抽词标引和多文档单主题赋词标引。其中，基于 DCF的文本聚类算法可归入多文档多主题自动抽词标引方法。

（2）文本聚类结果描述与自动摘要

自从 1958年 Luhn开始进行自动摘要研究开始，该研究就受到广泛关注[19]。自动摘要（Automatic Summarization）的目标对一个单文档或一组文档（多文档）进行主题概括，并给出简洁的文字描述[8]，相应的，自动摘要分为单文档摘要和多文档摘要。文本聚类结果描述与多文档摘要的类似之处在于，他们都需要给一堆文本进行综合和概括，并给出可读性强、连贯的、简洁的文字描述。关键词可以看成文摘句的特例。因此，多文档摘要上的一些方法可以应用到文本聚类结果描述上。此外，自动文摘研究中，文摘句的测评方法对文本聚类结果描述的测评也提供了参考作用。

文本聚类结果描述与自动摘要的不同之处在于，前者在生成聚类描述后，聚类描述和生成的每个类簇的文档是进行关联的，而自动摘要生成的仅为文摘句，不用将文摘句和每个文档进行关联，并且，文摘句的长度一般大于文本聚类结果描述词的长度[8]。

（3）文本聚类结果描述与概念聚类

概念聚类（Conceptual Clustering）是数据挖掘与机器学习中的一种聚类方法，给出一组未标记的对象，它产生对象的一个分类模式[1]。数据挖掘与机器学习中对聚类的要求之一：用户希望聚类结果是可解释的，可理解的和可用的。聚类需要和特定的语义解释和应用相联系[20]。概念聚类由两个部分组成，即先发现合适的簇，然后形成对每个簇的描述[10]。聚类质量不再只是单个对象的函数，而且加入了如导出概念描述的简洁性和一般性等因素[10]。自从上个世纪 80年代，Michalski首先提出概念聚类方法以来[20-21]，随后人们提出了大量的概念聚类算法。概念聚类的绝大多数方法采用统计学的方法，概率描述用于描述导出的概念[10]。COBWEB[22]就是一种流行的简单增量概念聚类算法，由 Fisher于 1987年提出。其他的概念聚类算法包括：CYRUS[23]、UNIMEM[24]、WITT[25]、LABYRINTH[26]、GALOIS[27]、DNF描述方法[28]、IITERATE[29]、GCF[30]、SUBDUE[31]等。

聚类描述与概念聚类的相似之处在于他们都要给类簇进行标注，表明类簇的主要内容。传统的概念聚

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

类算法通常使用特征的概念分布、约束特征的上下限等表达方法，用在文本聚类的结果描述上，仍然存在可读性不强和不易理解等问题。

通过上面的比较分析可以看出，与文本聚类结果描述任务最为接近的是自动标引，文本聚类结果描述可以简化为多文档单主题抽词标引。本文拟从多文档单主题抽取角度来解决文本聚类描述这个问题，下面将对该问题进行形式化。

3 聚类描述的要求

本节详细说明文本聚类描述所要达到的特殊要求。由于网页或文档目录描述款目实质上是一种类别描述形式，因此，根据网页或文档目录描述款目来研究文本聚类描述的一些基本特征，由此归纳出类别描述要达到的基本要求。

图 3 网页或文档目录

本文利用如图 3所示的（a) Google目录[32]、（b) Yahoo! 商业与经济类[32]、（c) Sogou工商经济类[34]以及（d) CNKI主题库 (足球类)[35]等四个网页或文档目录为分析对象。对这个目录描述词进行归并处理后，最后得到类别描述词 900余个，对他们进行了长度分布统计、短语类型统计分别如表 2和表 3所示。

表 2 类别描述词长度分页

长度	3	4	5	6	8	9	10	11	12	13	14	15	16	17	18	20
数量	1	271	1	103	306	29	71	10	40	15	31	9	8	2	1	4

由表 2可以看出，类别描述词以长度为 4个汉字、两个汉字以及三个汉字的词语或短语为主，这三种长度类型的词语的比率达 75. 56%，这与本文第 4部分中对关键词类型的分析结果是比较一致的。由表 3可以看出，类别描述词以单个词语（A)与两个词语组成的短语（A/B)为主，两者占总描述词数的 82. 44%。另外，每一种短语类型中，含有名词词性成分的短语类型占该类的比率分别为：91. 89%、80. 19%、70. 54%、81. 82%、71. 43%。由此可知，类别描述词主要以含有名词词性的词语或短语构成。因此，在进行类簇描述短语的提取时，对含有名词词性的词语或短语应赋较大的权重。

表 3 类别描述词短语类型分布

短语类型	A		A/B		A/B/C		A/B/C/D		A/B/C/D/E	
数量	530		212		129		22		7	
POS类型	N	~N	N	~N	N	~N	N	~N	N	~N
数量	487	43	170	42	91	38	18	4	5	2
比率（%）	91. 89	8. 11	80. 19	19. 81	70. 54	20. 46	81. 82	18. 18	71. 43	28. 57

(注：N表示含有名词词性成分的短语类型，包括：ng、n、nr、ns、nt、nz等，（~N)表示 N以外的词性。)

类簇描述词在长度和短语类型上的统计分布情况，可以用于类簇描述任务，通过加权的方式，对描述词作基本的长度和短语类型约束。除此之外，文本聚类对其聚类结果描述还有以下特殊要求。

（1）简洁性

简洁性（Concíseness)要求就是在能表达类簇信息的前提下，使聚类描述尽量简短[8]。描述词的长度是对描述词简洁性最简单的度量，可以利用描述词的字数或其包含的单词数来度量。Byron J Gao& Martin Ester在研究聚类描述格式问题时，提出利用最小描述长度原理（Minimum Description Length, MDL)和最大描述精确率原理（Maximum Description Accuracy, MDA)对聚类描述进行综合考虑的方法[36]。

（2）易理解性

易理解性（Understandability)，也可称为易读性。情报学、语言学、传播学和心理学等领域对易读性有专门研究。在情报语言学研究领域中，词类、词形、词义等词汇控制方法被用来对主题描述词进行控制[37]。语言学中文本易读性的定义为文本易于阅读和理解的程度或性质，其取决于多种因素，主要包括词长、不同词的比例、词汇的抽象程度、代词数、介词数、词缀数等，很多语言学研究者对文本的语言学特征进行量化分析，提出很多易读程度计算公式[38]。传播学中对易

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

读性的研究，关系到类别描述词的易读性的研究主要包括：对字词形式的约束（如多用常见字、尽可能选用实体动词和及物动词、尽量少用形容词和介词等），词汇迷雾指数，即词汇的抽象程度、艰涩程度等[39]。心理学领域通常使用回答问题法来进行易读性测试，即：读者在读完文本后回答问题，根据回答问题的正确度来判断文本的易读度[38]。

需要指出的是，文本聚类结果描述对易读性的要求与上面所提到的三个领域中研究的易读性不同的是，作为聚类结果描述的描述词，除了词语本身的理解性之外，还包括通过词语映射类簇中所包含文本的能力。DawidWeiss在其论文中将聚类描述的这种性质称为“透明度”（Transparency）[8]，而 KrishnaKummamuru等人则将其称为标签的预测力（Predictiveness）[40]。

（3）精确性

精确性（Accuracy）要求描述词能反应所对应类簇的主题内容。不同于传统的主题标引，聚类描述的对象不再是单个文本，而是含相同主题的文本类簇。

一般通过度量描述词与类簇之间的相符度来度量描述词的精确度，可以借用分类中词语与类别之间的关联度来解决这个问题。1997年，YimingYang& JPedersen利用 χ^2 来度量词语与类别之间的关联度[41]。2006年 Yuen-HsienTseng& Chi-JenLin等人则采用词语与类别的相关系数及其变种形式进行描述词与类别之间的关联度度量[9]。本文拟采用简化的方法来进行描述词的精确性度量，即利用描述词在类簇中的文档频次进行度量。

（4）区分性

描述词必须具有一定的区分性（Distinctiveness），即要求描述词在其描述的类簇中频繁出现，而在其他类簇中很少出现。HananAyad& MohamedKamel在进行话题发现研究时，利用一种变种的 TF×IDF方法计算描述词的权重[42]，该公式如式（1）所示。

$$w_{ij}=tf_{ij}\log\frac{K}{cf_j} \qquad (1)$$

其中，w_{ij} 为类簇 i 中的描述词 j 的权重；tf_{ij} 是类簇 i 中的描述词 j 的出现频次，即：类簇 i 中出现描述词 j 的文档数；$\log\frac{K}{cf_j}$ 为逆类簇频率（InverseClusterFrequency，ICF），其中 K 为类簇总数，cf_j 为类簇频率，即：描述词 j 出现的类簇的总数。

此外，PuckradaTreeratpituk& JamieCallan在研究文本聚类的类簇标注方法时，也提出与 TF×IDF方法类似的度量描述词的区分性的计算方法[2]。

需要指出的是，精确性和区分性是两个比较难以区分的特性，例如 χ^2，相关系数本身也具有一定区分功能。

4　聚类描述问题的形式化

给定文档集合 D，通过聚类算法得到 D 的一个划分为 $C=\{C_i|C_i\subseteq D\}$，$D=\bigcup_{C_i\in C}C_i$，$i\in[1,K]$，K 为类簇总数，且 $\forall C_i,C_j(i\neq j,i,j\in[1,K])$，$\exists C_i\neq C_j$。

一般地，聚类描述是从候选的描述词中选择最佳的词语作为类簇描述词。本文首先给出候选主题模式的定义。

定义 1：候选主题模式（TopicPatternCandidate，TPC）。给定一个类簇 $C_i(i\in[1,K])$ 的主题描述词（Topic）集合 $T_i=\{t_1,t_2,\cdots,t_n\}$，一个主题描述词 $t_{ij}(j\in[1,n])$，给定常数 σ、ε、Max_LEN，若 t_{ij} 满足如下基本约束条件：

①长度约束：$LEN(t_{ij})\leqslant Max\_LEN$，$LEN(t_{ij})$ 表示 t_{ij} 的词长；

②词性约束：$POS(t_{ij})\cap N=\Phi$；$POS(t_{ij})$ 表示 t_{ij} 的词性，N 的含义参见表 3；

③精确度约束：$DF(t_{ij})>\sigma$，$DF(t_{ij})$ 表示 t_{ij} 在 C_i 内的文档频率；

④区分性约束：$ICF(t_{ij})>\varepsilon$；

则称 t_{ij} 为类簇 i 的一个候选主题模式，记为 tpc_{ij}。

由于一篇文档可能涉及多个主题，多个文档更加有可能关联到多个主题，因此一个类簇可能有多个候选主题模式。若把多个候选主题模式以词语列表的形式返回给用户，在多个主题模式缺乏足够语义关系的情况下，会增加聚类描述可理解性上的困难。而词语列表形式本身就需要用户去进行描述词之间的语义推理，判断类簇的主题，这样会增加用户的智力负担。因此，聚类描述的最终目标是要从候选主题模式中寻找主题模式，即一个最能概括类簇主题内容的词语或短语。

本文将文本聚类结果描述，即每个类簇的标注（ClusterLabels）定义为主题模式，给出定义如下。

定义 2：主题模式（TopicPattern，TP）。给定类簇 i 的候选主题模式 $tpc_{ij}(i\in[1,K],j\in[1,n])$，利用函数 $\tau(\cdot)$ 对 tpc_{ij} 进行重要性排序，得分最高的候选描述词，即 $Max(\tau(tpc_{ij}))$ 被称为主题模式，记为 tp_i。

其中，函数 $\tau(\cdot)$ 为度量候选主题重要性的函数，

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

本文在后面将利用 SVM、多元线性回归模型、Logistic 回归模型等三个统计机器学习模型和一个基准模型来度量候选主题的重要性，最终确定类簇的描述词。

根据定义 2，可以将聚类描述（Cluster Description）这一任务描述如下。

聚类描述是指对文本聚类生成的每个类簇，通过重要度、区分性、长度、词性等基本约束条件生成的候选主题模式，然后通过候选主题模式重要性度量函数对候选主题模式进行得分评估，将每个类簇中得分最高的候选主题模式作为该类簇的类簇描述的过程。

聚类描述任务的形式化定义如下。

①文档集合 D，通过聚类算法得到 D的一个划分为 $C=\{C_i \mid C_i \subseteq D\}$，$D=\bigcup_{C_i \in C} C_i$，$i \in [1, K]$，K为类簇总数，且 $\forall C_i, C_j (i \neq j, i, j \in [1, K])$，$\exists C_i \neq C_j$；

②每个类簇 C_i（$i \in [1, K]$）的描述词集合 $W_i=\bigcup_{d \in C_i} W_d$，若通过函数 $\tau(\cdot)$可以建立 W到 C_i（$i \in [1, K]$）的映射，即：$\tau(W_i) \to C_i$，$i \in [1, K]$；则称函数 $\tau(\cdot)$为聚类描述函数。

5 聚类描述评价方法

文本聚类结果描述可简化为多文档单主题自动标引。因此，聚类描述的评价方法可以借鉴自动标引评价方法。传统自动标引研究具有近 50年的历史，相应的出现了很多的评价方法，而文本聚类描述作为最近才引起人们注意的一个研究话题，除了在聚类描述方法较少外，对聚类描述的评价方法也相对较少。目前所使用的聚类描述评价方法主要有标引结果比较法与用户可接受性评价两种方法。

（1）标引结果比较法

设测试集中的类簇描述词总数为 n，则聚类描述的结果可以表示如表 4所示。本实验将人工标引的结果分为两种情况，即：人工标引为描述词的情况与人工标引为非描述词的情况，后者就是将人工标引描述词后，类簇描述候选词语剩下的词作为非描述词。

表 4 聚类描述结果评价列联表

	人工标引为描述词	人工标引为非描述词
系统标引为描述词	a	b
系统标引为非描述词	c	d

其中，$n=a+b+c+d=n_t+n_f$，$n_t=a+c$为人工标引的描述词总数；$n_f=b+d$为人工标引为非描述词的总数。

①查准率（Precision）

$$P=\frac{a}{a+b} \tag{2}$$

查准率 P即描述词的标对率，查准率反映了聚类描述系统找对描述词的能力，查准率越大，将描述词误判为非描述词的可能性越小。

②召回率（Recall）

$$R=\frac{a}{a+c} \tag{3}$$

召回率 R即描述词的检出率，它反映了聚类描述系统发现描述词的能力，召回率越高，无法标引描述词的情况就越少。

③ F_1值

$$F_1(P,R)=\frac{2PR}{P+R} \tag{4}$$

F_1测度是关于查准率与召回率的调和平均值。

2006年，Pucktada Treeratpituk & Jamie Callan在进行聚类描述的测评时，将描述词与参照的类别标签的匹配分为精确匹配和部分匹配两种情况，进行了描述词的 Match@N、Top—N查准率（P@N）、MMR、MTRR等指标测试[2]。

（2）用户可接受性评价

与自动标引评价中的用户可接受性评价方法类似，聚类描述的用户可接受性评价是根据聚类描述的可读性，人工进行打分。2004年，Krishna K.和 Rohit K.等人在进行聚类结果聚类评价时涉及到聚类描述的评价，使用的方法是根据事先设计好的问题域，让志愿者对几种描述方法生成的结果进行两两比较，最终统计确定最优的聚类方法[40]。2006年，Hua Li和 Dou Shen等人在进行聚类描述评价时采用了人工打分的方法，即根据聚类描述的可读性进行三分制的打分[7]。同年，Yuen—Hsien Tseng等人则是人工直接判断几种方法生成的聚类描述，被评为“最好”的次数最多的方法就是最优的聚类描述方法[9]。

6 结 语

本文首先对文本聚类结果描述的研究背景和相关的研究情况进行说明，分析文本聚类结果描述的相关研究工作与相关研究领域，定位文本聚类结果描述研究内容；分析文本聚类结果描述的具体要求，并对聚类结果描述问题进行形式化。最后给出文本聚类结果描述的评价方法。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目前解决文本聚类描述任务的主要方法为统计方法，借助于外部资源可以在一定程度上提高描述的质量。作者在文献[43]中利用统计机器学习模型对聚类描述进行研究，结果表明，SVM模型在解决该问题时性能是最优的。通过实验结果还表明，聚类描述任务呈一定的线性趋势。另一方面，基于机器学习的文本聚类描述算法还存在诸如训练数据集标注问题、识别和获取更加有效的聚类描述特征等问题。

参考文献：

[1] Popescul A, Ungar L. Automatic Labeling of Document Clusters [EB/OL]. [2007－01－10]. http://www.cis.upenn.edu/~popescul/Publications/popescul00labeling.pdf

[2] Pucktada T, Jamie C. Automatically Labeling Hierarchical Clusters [C]. In Proceedings of the 2006 International Conference on Digital government research. San Diego, CA, USA: 2006: 167－176

[3] Maqbool O, Babri H A. Interpreting Clustering Results through Cluster Labeling [C]. In Proceedings of the IEEE International Conference on Emerging Technologies (ICET'05), Islamabad, Pakistan: 2005: 429－434

[4] Stein B, Meyer zu Eissen S. Topic Identification: Framework and Application [C]. In Proceedings of the 4th International Conference on Knowledge Management (I-KNOW'04), Graz, Austria: 2004: 353－360

[5] Lawrie D, Croft W B, Rosenberg A L. Finding Topic Words for Hierarchical Summarization [C]. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), New Orlean, LA, USA: 2001: 249－357

[6] Muscat R. Automatic Document Clustering Using Topic Analysis [R]. Technical Report CSAI2005－01, Department of Computer Science & AI, University of Malta: 2005: 1－16.

[7] Li H, Shen D, Zhang B Y, et al. Adding Semantics to Email Clustering [C]. In Proceedings of the IEEE 6th International Conference on Data Mining (ICDM'06). Hong Kong, China: 2006: 18－22.

[8] Dawid W. Descriptive Clustering as a Method for Exploring Text Collections [D]. Poznan University of Technology, Poznań, Poland: 2006: 7－56.

[9] Tseng Y H, Lin C J, Chen H H, et al. Toward Generic Title Generation for Clustered Documents [C]. In Proceedings of the 3rd Asia Information Retrieval Symposium (AIRS2006), Singapore: 2006: 145－157.

[10] Han J, Kamber M. Data Mining: Concepts and Techniques [M]. San Francisco: Morgan Kaufmann: 2001: 376－379.

[11] Glenisson P, Glänzel W, Janssens F, et al. Combining Full Text and Bibliometric Information in Mapping Scientific Disciplines [J]. Information Processing & Management: 2005, 41(6): 1548－1572

[12] Lai K K, Wu S J. Using the Patent Co-citation Approach to Establish a New Patent Classification System [J]. Information Processing & Management: 2005, 41(2): 313－330

[13] Cutting D R, Karger D R, Pedersen J O, et al. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections [C]. In Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), Copenhagen, Denmark: 1992: 318－329.

[14] Cutting D R, Karger D R, Pedersen J O. Constant Interaction-time Scatter/Gather Browsing of Large Document Collections [C]. In Proceedings of the 16th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93), Pittsburgh PN, USA: 1993: 126－135.

[15] Muller A, Dorre J, Gerstl P, et al. The TaxGen Framework: Automating the Generation of a Taxonomy for a Large Document Collection [C]. In Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS'1999), Maui, HI, USA: 1999: 2034－2042.

[16] Anton V L, Croft W B. An Evaluation of Techniques for Clustering Search Results [R]. Technical Report IR－76, Department of Computer Science, University of Massachusetts, Amherst: 1996: 1－19.

[17] Zamir O, Etzioni O. Web Document Clustering: A Feasibility Demonstration [C]. In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), Melbourne, Australia: 1998: 46－54.

[18] Glover E, Pennock D M, Lawrence S, et al. Inferring Hierarchical Descriptions [C]. In Proceedings of the 11th International Conference on Information and Knowledge Management (CKIM2002), McLean, VA: 2002: 4－9.

[19] Luhn H P. The Automatic Creation of Literature Abstracts [J]. IBM Journal of Research and Development: 1958, 2(2): 159－165.

[20] Michalski R S, Stepp R E. Learning from Observation: Conceptual Clustering [A]//Michalski R S, Carbonell J G, Mitchell T M eds. Machine Learning: An Artificial Intelligence Approach [C], San Mateo, CA: Morgan Kauffmann: 1983: 331－363.

[21] Michalski R S. Knowledge Acquisition through Conceptual Clustering: A Theoretical Framework and an Algorithm for Partitioning Data into Conjunctive Concepts [J]. Journal of Policy Analysis and Information Systems: 1980, 4(3): 219－244.

[22] Fisher D H. Knowledge Acquisition via Incremental Conceptual Clustering [J]. Machine Learning: 1987, 2: 139－172

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[23] Kolodner J L. Reconstructive Memory: A Computer Model[J]. Cognitive Science, 1983, 7: 281-328.

[24] Lebowitz M. Experiments with Incremental Concept Formation[J]. Machine Learning, 1987, 2: 103-138.

[25] Hanson S J, Bauer M. Conceptual Clustering, Categorization and Polymorphy[J]. Machine Learning, 1989, 3: 343-372.

[26] Thompson K, Langley P. Incremental Concept Formation with Composite Objects[C]. In: Proceedings of the 6th International Workshop on Machine Learning (ICML-89), Ithaca, NY, USA, 1989: 373-374.

[27] Carpineto C, Romano G G. An Order-theoretic Approach to Conceptual Clustering[C]. In: Proceedings of 10th International Conference on Machine Learning, Amherst (ICML-93), MA, USA, 1993: 33-40.

[28] Agrawal R, Gehrke J E, Gunopulos D, et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications[C]. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD98), Seattle WA, USA, 1998: 94-105.

[29] Biswas G, Weinberg J B, Fisher D H. Iterate: A Conceptual Clustering Algorithm for Data Mining[J]. IEEE Transactions on Systems, Man, and Cybernetics (Part C), 1998, 28(2): 100-111.

[30] Talavera L, Béjar J. Generality-based Conceptual Clustering with Probabilistic Concepts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23: 196-206.

[31] Jonyer I, Cook D J, Holder L B. Graph-based Hierarchical Conceptual Clustering[J]. Journal of Machine Learning Research, 2001, 2: 19-43.

[32] Google网页目录[EB/OL]. [2007-02-01]. http://www.google.com/dirhp?hl=zh-CN.

[33] Yahoo! Business and Economy[EB/OL]. [2007-02-01]. http://gb.chinese.yahoo.com/Business_and_Economy/.

[34] 工商经济. 搜狐分类目录[EB/OL]. [2007-02-01]. http://www.sogou.com/q002/q002.html.

[35] CNKI主题数字图书馆[EB/OL]. [2007-02-01]. http://topic.cnki.net/search.aspx?class=a1.

[36] Gao B J, Ester M. Cluster Description Formats, Problems and Algorithms[C]. In: Proceedings of the Sixth SIAM International Conference on Data Mining (SDM06), Bethesda, MD, USA, 2006.

[37] 侯汉清, 马张华. 主题法导论[M]. 北京: 北京大学出版社, 1991: 16-18.

[38] 晏生宏, 黄莉. 英文易读度测量程序开发探索[J]. 重庆大学学报(社会科学版), 2005, 11(2): 92-97.

[39] 邵培仁. 传播学[M]. 北京: 高等教育出版社, 2000: 131-132.

[40] Kummamuru K, Lotlikar R, Roy S, et al. A Hierarchical Monothetic Document Clustering Algorithm for Summarization and Browsing Search Results[C]. In: Proceedings International WWW Conference (WWW2004), New York, NY, USA, 2004: 658-665.

[41] Yang Y M, Pedersen J. A Comparative Study on Feature Selection in Text Categorization[C]. In: Proceedings of the International Conference on Machine Leaning (ICML'97), Nashville TN, USA, 1997: 412-420.

[42] Ayad H, Kamel M. Topic Discovery from Text Using Aggregation of Different Clustering Methods[C]. In: Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, 2002: 161-175.

[43] 章成志. 主题聚类及其应用研究[D]. 南京: 南京大学, 2007: 28-50.

(作者 E-mail: zhangchz@istic.ac.cn)

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net