主题聚类及其应用研究

國家圖書館出版社

主题聚类及其应用研究

TOPIC CLUSTERING AND ITS APPLICATIONS

章成志 著

國家圖書館出版社

图书在版编目(CIP)数据

主题聚类及其应用研究 / 章成志著. —北京:国家图书馆出版社,2013.4
ISBN 978-7-5013-4929-6

Ⅰ.①主… Ⅱ.①章… Ⅲ.①主题—情报检索—聚类分析 ②自然语言处理—自动标引 Ⅳ.①G354 ②TP391.1

中国版本图书馆 CIP 数据核字(2013)第 026908 号

责任编辑:金丽萍

书名 主题聚类及其应用研究
著者 章成志 著

出版 国家图书馆出版社(100034 北京市西城区文津街 7 号)
(原北京图书馆出版社)
发行 010-66114536 66126153 66151313 66175620
66121706(传真) 66126156(门市部)
E-mail btsfxb@nlc.gov.cn(邮购)
Website www.nlcpress.com→投稿中心
经销 新华书店
印刷 北京科信印刷有限公司

开本 787×1092(毫米) 1/16
印张 11.75
版次 2013 年 4 月第 1 版 2013 年 4 月第 1 次印刷
字数 250(千字)

书号 ISBN 978-7-5013-4929-6
定价 48.00 元

本书出版获得以下研究项目的资助：

- 国家自然科学基金青年项目（70903032）：基于可比语料的多语言文本聚类研究
- 中央高校基本科研业务专项资金资助项目（NUST2011ZDJH15）：Web 2.0 环境下多语言标签自动聚类研究

序

一直以来，信息组织研究就是图书情报学科重要的研究内容。信息组织方法是针对信息检索的需要，对信息资源进行内容分析、标引、处理，最终使得信息资源有序化的方法。信息组织方法在互联网的应用服务中发挥巨大作用。当前信息组织方法存在的问题是，一方面由于主题法、分类法或分类主题一体化方法依赖于大量的人工参与，使得绝大多数机构面临人力、物力和财力资源不足的困境。传统信息组织方法置身于互联网海量数据的环境中，无法充分、及时地为满足用户信息需求提供便利。另一方面，数据挖掘、机器学习等人工智能技术在互联网应用服务中发挥人工无法替代的作用的同时，由于存在高维数据计算问题并且缺乏充分的主题控制和语义理解机制，导致应用服务过程中难以及时进行响应，出现大量的信息噪声，从而影响服务质量。

为了有效解决以上两个问题，章成志博士将信息组织方法中的主题方法与数据挖掘、机器学习中的聚类方法相结合，从主题角度出发，提出了主题聚类方法，指出主题聚类中存在的5个方面的问题：如何增强主题提取评估的可靠性并降低主题提取评估成本？如何提高主题提取的实用性？如何提高聚类对象相似度计算的可靠性？如何提高基于样本加权的文本聚类方法的实用性？如何增强文本聚类结果的可读性？

本书就上述5个方面的问题进行了深入研究，其研究具有重要的理论创新与实际应用意义。我认为本书的价值主要体现在如下两个方面。

其一，对自动标引工作进行了全面总结，并对以往工作进行了改进。具体来说包括如下几个方面的创新性工作：

（1）提出自动标引的通用评价模型。针对常规自动标引评价方法存在的评价结果不能完全反映真实标引结果以及评价成本过高的情况，作者提出一种通用的自动标引评价模型，该模型有效利用外部资源，根据有参照情况与无参照情况，分别对标引结果进行评价，增加评价的可靠性并降低评价的成本。

（2）提出基于机器学习的关键词自动提取算法。为了有效利用标引对象的特征，并考虑到标引可以转换为序列标注问题，作者利用条件随机场模型进行关键词的自动提取研究；融合多个标引模型的标引结果进行投票学习，提出基于集成学习策略的自动标引方法。实验结果表明该方法在一定程度上能改善自动标引的性能。作者还提出基于Citation-KNN的自动赋词标引算法，提高赋词标引的实用化程度。

其二，对文本聚类中的相似度计算、聚类样本权重对聚类的影响、聚类描述等关键问题进行了创新性研究。

（1）提出基于多层特征与基于多语境的聚类对象相似度计算方法。针对计算字符串相似度传统方法的不足之处，作者提出以相似元作为字符串的基本处理单元，综合考虑相似元的字面、语义及统计关联等多层特征的字符串相似度计算方法。实验结果表明该算法的有效性。通常，某一查询式在不同的语境下，从不同侧面反映了该查询式的语义，作者利用语料库、释义词典、用户搜索日志作为查询式的不同的语境，进行基于多语境的查询式相似度计算方法，并

将该算法用于查询词的相关词的自动获取应用中。

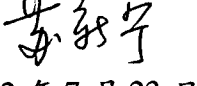
(2)提出基于样本加权的文本聚类算法。作为一种逐渐引起人们注意的算法,样本加权聚类算法还存在一些需要解决的问题,例如聚类对象之间的结构信息对样本加权聚类是否有帮助,如何将结构信息自动转换为样本或对象的权重?针对该问题,作者以学术论文为聚类对象,以 K-Means、Fuzzy C-Means 算法为聚类算法基础,利用论文之间的引用关系计算每篇论文的 PageRank 值,并将其作为权重,提出一种基于样本加权的新的文本聚类算法。实验结果表明,基于论文 PageRank 值加权的聚类算法能改善文本聚类效果。作者利用该算法进行两个方面的应用,即基于主题聚类的主题数字图书馆的设计与实现,基于主题聚类的学科热点的检测。

(3)提出基于机器学习的文本聚类结果的描述算法。标注文档集合聚类后生成的类簇,是主题聚类应用中一项重要并富有挑战性的任务。针对文本聚类结果可读性较弱问题,作者提出了一种增强聚类结果的可理解性与可读性的算法,即基于支持向量机等机器学习方法的文本聚类结果描述算法。为了进一步提高类簇描述词的质量,作者提出了一种基于 DCF-DCL 组合策略的文本聚类结果描述算法。实验结果表明这两个算法所取得的效果要优于常规的聚类结果描述方法。作者综合利用主题提取、文本聚类、聚类描述等方法用于搜索结果聚类中。

当前,已经有图书情报、计算机、新闻传播等多个领域的研究者对互联网海量信息的组织与利用进行深入研究。在图书情报、计算机、语言学等多个学科有效融合的基础上,信息组织的研究呈现出新的生机与活力。章成志博士近十年来,一直致力于信息组织相关研究,努力吸收计算机与语言学学科中的相关理论与方法,解决图书情报领域中的信息组织问题。本书正是在文本自动标引的基础上,进行基于主题的文本聚类研究,创新性地提出主题聚类中面临的问题并给出相应的解决方案。

章成志勤于思考,并且潜心向学。在信息组织理论与方法上进行研究的同时,也注意研究对象的拓展。随着多语言、社会化的信息在增长,对多语言、社会化的信息进行有效组织更是具有挑战性的问题。章成志结合这一特色,并凭借其坚实的研究积累,对多语言文本聚类、基于社会化信息的文本聚类等亦开展了相关的研究,体现了他对学科前瞻性研究具有很好的预测能力。

我是章成志同志的博士生导师。我很欣慰地看到章成志能将各个阶段的研究成果进行总结,分享给同行。该书是他在主题聚类方面研究的工作总结。这本书内容丰富、数据翔实、论证充分,是图书情报、计算机与语言学等学科交叉研究的代表作之一。衷心期望该书的出版能够进一步推动信息组织研究的深入与拓展,同时祝愿章博士的学术研究之路越走越宽。



2012 年 7 月 23 日

于南京大学

目　录

第1章 引言

1.1 研究背景

我们正处于“信息爆炸”的时代。为什么当各类信息像洪水一样向我们涌来时，我们仍然缺乏所需要的信息呢？这是因为在信息社会之中，“没有控制和没有组织的信息不再是一种资源。它倒反而成为信息工作者的敌人”。[1]与此同时，信息需求用户面临“可以选择的信息越多，便越难作出选择”这样的窘境。因特网是最主要的信息源，然而，其组织和使用技术的发展往往跟不上因特网信息的增长。全文数据库、搜索引擎等内容服务可以为人们查找与关键词相关的文档，但返回的结果往往是文档数量太多，并且命中率不高。出现该问题的根本原因是以关键词为基础的传统检索手段无法真正满足用户的检索需求，用户在构造查询式、浏览查询结果方面的时间与智力开销较大。

全文数据库、搜索引擎等信息服务业务成功的关键之一就在于提高信息服务的质量，即如何最大限度地推测用户的查询意图、将相关程度高的信息以合理的方式提供给用户，使得用户能在较短时间内得到其需要的信息。用户迫切希望的信息服务是提供信息噪声尽量少乃至没有噪声的服务。当前，信息组织方法与数据挖掘、机器学习技术在提高信息服务质量上发挥越来越大的作用。

一方面，信息组织方法在有效开发利用信息资源方面发挥重要作用。其中，主题法作为一种从内容角度标引和检索信息资源的信息组织方法，[2]可以有效弥补全文检索的不足。主题法按选词方式的不同，可以分为标题法、元词法、叙词法和关键词法等。[3][4]标题法和元词法没有真正从文本内容或概念上对文本进行组织，已经逐步被叙词法和关键词法所代替。叙词法以规范化的自然语言做标识，主要以标识的概念组配来对信息主题进行描述，但叙词表的编制和管理难度大，标引难度大。关键词法直接利用自然语言中未经控制或只做少量控制的词语表达主题概念，易于实现自动标引。[4]现阶段，由于主题法、分类法或分类主题一体化方法依赖于大量的人工参与，使得绝大多数机构面临人力、物力和财力资源不足的困境。传统信息组织方法置身于互联网海量数据的环境中，无法充分、及时地为满足用户信息需求提供便利。

另一方面，由于人工方法在组织管理因特网海量信息时存在困境，迫使人们去寻找解决海量数据自动化处理的方法和技术。大量研究表明，自动标引[5][6]、自动摘要[7]、自动分类[8]、自动聚类[9]等文本挖掘方法和技术可以有效地组织、管理和分析大规模文本信息资源，向用户提供高效的信息服务。怎样从海量的文本数据中抽取和发掘有用的信息和知识已成为一个日趋重要的问题。由于这个原因，文本挖掘虽是一个新兴学科，但已成为一个引人瞩目、发展迅速的领域。[10]从机器学习角度出发，文本挖掘的方法可以分为监督学习和非监督学习的两种方法。基于监督学习的文本挖掘需要大量人工标注的语料进行训练，这种方法面对更新频繁、实时性较强的网络信息适应性较差。文本聚类是一种典型的非监督学习方法，信息检索中的聚类没有分类词表、训练集或决策树做依据，完全利用信息相似原理来进行聚类，即利用信息之

间标引词的相似度来确定两条信息是否成为一类。[11]然而，由于一般的文本聚类方法直接以词语作为特征，经过停用词过滤等预处理后，文本向量空间的维度可能会达到数万以上，这样使得应用服务过程中面临着高维数据计算问题，导致服务难以及时响应。同时，由于缺乏充分的主题控制和语义理解机制，导致聚类等文本挖掘技术在应用服务过程中出现大量的信息噪声，影响服务质量。语义网的理论与应用的目标就是解决当前互联网缺乏语义理解这一问题。作为语义网理论与应用中的基础性工作，本体构建本身也同样面临着传统信息组织方法中存在的问题。

针对传统信息组织方法存在的困境、一般文本挖掘方法存在高维数据计算并缺乏主题控制问题以及互联网应用服务质量提升的迫切需求，信息组织方法与机器学习方法的有效融合已是当务之急。信息组织方法中的主题法与数据挖掘、机器学习中的聚类方法的结合，使得主题聚类方法应运而生。主题提取是一项基础性的信息提取工作，主题聚类则是以主题提取为前提的信息聚类过程。提高主题提取、主题聚类质量与实用化程度是当前迫切需要解决的问题。

与传统的主题法、分类法、主题分类一体化以及索引法的信息组织方法和一般的聚类分析方法不同，主题聚类主要通过对文本、查询式等聚类对象进行基于机器学习的主题分析，将聚类对象转换为基于主题的表示形式，以达到降低特征空间维度的目的，然后以主题表示为基础进行对象的聚类分析，最后得到基于主题的聚类结果描述。

如图 1 - 1 所示，主题聚类（或者称为主题聚类一体化）信息组织方法，正是融合信息组织方法中的主题法与数据挖掘、机器学习中的聚类方法形成的一种新的信息组织方法。

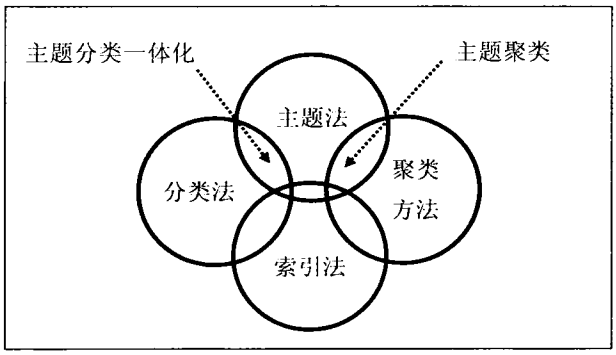


图 1 - 1　信息组织方法关系图

从主题聚类的过程可以看出，主题聚类方法具有如下 3 个方面的优势。

首先，主题聚类以主题分析、主题提取和描述为基础，可以发挥主题法在组织信息方面的优势，对聚类特征进行主题或语义控制，提高信息服务的质量。

其次，主题聚类是在聚类对象的主题提取基础上进行的，通过主题提取可以对聚类对象进行维度约简，从而避免高维数据计算问题，大大缩短信息服务的响应时间。

最后，主题聚类方法不同于传统的文本聚类方法在于它还可以对聚类的结果进行基于主题的描述，提高聚类结果的可读性与可理解性。

目前信息组织中的主题法、分类法、主题分类一体化以及索引法的研究已经非常深入，同

时，数据挖掘与机器学习中的聚类分析方法与技术的研究也如火如荼，但对主题法与聚类方法的融合体，即主题聚类的研究，则很少有人涉及。近年来，有个别学者对主题聚类进行了相关研究。2003 年、2005 年，Kang、Chang 与 Hsu 等分别进行基于关键词聚类的文本聚类研究。[12][13]2003 年，孙学刚和陈群秀等人采用二次特征提取和聚类方法，结合密度算法和 K 近邻准则对 Web 文档按照主题进行聚类。[14]2005 年，Zhao 和 George Karypis 提出主题驱动的文本聚类方法，将文本聚类问题转换为半监督学习问题。[15]2006 年，马张华和陈文广等人提出基于控制词集的中文信息动态自动聚类技术，其中，控制词集由关键词表、等同控制词表、等级词表等词表组成，并被用作文本切分工具、聚类依据、聚类结果优化以及类别描述。[16]2007 年，赵世奇和刘挺等人提出一种基于主题的文本聚类方法，通过整合语言学特征得到文本的主题元素，并按主题元素索引完成聚类。[17]值得注意的是，这些研究基本上只涉及主题聚类研究的一个部分，除了缺少对主题聚类进行系统化的研究之外，还缺乏对主题提取质量的分析与聚类结果描述方法等方面的深入研究。

本书正是基于信息组织方法理论的探索与创新、信息服务质量亟须提升的迫切需求，提出主题聚类的方法，并针对主题聚类研究现状提出主题聚类中面临的问题，给出相应的解决方法，并将部分研究成果应用于实践。

1.2 研究意义

如上所述，正是因为信息组织方法理论上的探索与创新的需要、信息服务质量亟须提升的迫切需求，主题聚类研究才具有非常重要的理论意义和现实价值。本书综合情报语言学、文本挖掘、机器学习、自然语言处理、信息检索等多方面的理论与技术，提出基于主题聚类的基本思想和主要方法，并将主题聚类方法应用于主题数字图书馆构建、学科热点检测以及搜索结果聚类等信息服务中，具有重要的理论创新与实际应用意义，主要表现在如下两个方面：

1. 主题聚类方法是传统信息组织方法理论拓展的需要

如前所述，主题聚类与传统的主题法、分类法、主题分类一体化以及索引法的信息组织方法和一般的聚类分析方法不同，主题聚类主要通过对文本、查询式等信息集合进行主题分析，获得能表达主题的关键词或者主题词集合后，再对集合进行聚类分析，最后得到基于主题的聚类结果描述。主题聚类是主题聚类一体化的信息组织方法，它是融合信息组织方法中的主题法与数据挖掘、机器学习中的聚类方法形成的一种新的信息组织方法。

目前对主题法、分类法、主题分类一体化以及索引法等信息组织与数据挖掘、机器学习中的聚类分析方法的研究已经非常深入，但对主题聚类的研究则几乎无人涉及。正是基于对传统信息组织方法理论上拓展的需要，本书提出主题聚类的信息组织方法。主题聚类方法的提出具有重要的理论创新意义。

2. 主题聚类是信息服务质量提升的需要

信息服务质量的提升，迫切要求将主题聚类方法用于信息的组织与服务中。主题聚类可以解决海量信息组织与检索问题、解决查询式信息表示问题、解决查询结果组织问题以及其他相关信息服务问题。

（1）主题聚类方法能够解决海量信息组织与检索问题

传统信息组织方法主要有主题法、分类法、索引法。其中主题法用于主题标引和主题词查询,分类法用于分类标引、资源导航等,索引法用于信息资源的定位、使用户能准确地找出其所需信息。众所周知,这些传统的信息组织方法存在两个问题:一是这些信息组织方法要花费大量的人力、物力;二是它们忽视了聚类方法的价值,聚类可以发现知识。

从文本聚类的角度来看,通过对文本的主题分析与提取,可以对文本向量空间进行降维。Anton V. Leouski 和 W. Bruce Croft 的研究结果表明在文本聚类任务中,用 50—100 个高频词语表示文档内容已经足够,如果增加词的个数,反而降低系统性能。[9] 因此主题聚类可以避免聚类过程中的高维数据计算问题,缩短信息服务的响应时间。

从信息检索用户体验和信息检索系统可用性等方面考虑,在海量信息的形势下,当前的信息组织与检索方式并不能令用户满意。调查显示,绝大多数用户输入的中文查询词的长度不超过 3 个词,中文查询词长度为 1.85 个词[18],英文为 2.35 个词[19]。在海量数据集下,返回给用户的是成千上万条结果,使得用户无法立即获得其真正需要的信息。用户迫切希望通过信息检索系统能更加友好、快速地获得其需要的信息。传统的方法,如主题法、分类法、索引法在这方面显得无能为力,通过对包括查询式在内的海量数据进行主题聚类正是解决这一问题比较有前景的方法。

从检索的返回结果来看,传统结果表现形式一般是排序列表(Ranking list)方式,返回的结果太多,用户无法立刻找到其感兴趣的信息,并且用户一般并没有耐心浏览大量的返回结果。调查显示,搜索引擎用户中有 85% 的查询用户只翻看搜索引擎返回结果的第一个页面。[19] 此外由于查询结果本身可能包含多个主题(子主题,Subtopic),必须对查询结果进行主题划分。通过主题聚类方法可以解决查询结果的组织进行优化,便于用户进行信息获取。另外,通过对用户的聚类浏览行为还可以从一定程度上发现用户信息需求的兴趣偏好,根据这些浏览信息可以进行用户个性化信息服务。

(2)主题聚类方法能够解决查询式信息表示问题

目前信息检索模型主要有布尔模型、向量模型和概率模型 3 种经典模型,这些检索模型的查询式一般都由词汇构成。由词汇构造查询式在信息检索中存在 6 个深层次的问题:查询式构造难问题,也称查询式的"忠实表达"问题;同义词问题,也称查询词的"表达差异"问题;语义孤立问题,也称"词汇孤岛"问题;多义词问题,也称查询词"语境缺失"问题;歧义词问题;词语表达能力差异问题。

查询式表示问题对信息检索系统产生重要的影响。通过主题聚类方法,进行基于不同颗粒度的、不同语境下的查询式聚类,为用户提供与原查询式潜在相关的查询式,最大限度地解决查询式的表示问题。

(3)主题聚类能够提高其他相关信息服务的质量

学科热点检测从科技文献角度为学科的科学研究及管理提供动态、可辅助决策的学科发展状态信息,提高现有科技文献服务系统能力,提升科技文献信息服务水平。本书将主题聚类方法用于学科热点的检测中。根据大规模学科信息资源,进行主题提取和聚类,检测学科研究热点和话题,预测学科热点问题的研究趋势,为学科研究者和管理者提供借鉴。

综上可以看出,主题聚类的研究具有重要的理论与实际意义。本书侧重于对主题聚类研究中存在的一些问题进行分析,并给出解决这些问题的方法。

1.3 主题聚类研究中存在的问题与解决方法

如图 1 -2 所示,在信息检索过程中,涉及用户空间(包括用户信息需求和兴趣偏好等信息)、查询空间(包括用户查询行为的反馈信息)、文档空间(包括文档的链接与被链接信息)以及作者空间(包括信息源、责任者等信息)。对这些信息空间进行主题聚类可以提升信息服务的质量。由于在文本聚类领域研究成果较多,而在用户聚类、查询式聚类和返回结果聚类方面的研究成果不多,特别是在中文领域,目前仅极少数人在进行相关研究。根据文献调研,目前还暂时无人利用主题聚类方法对信息服务的各个阶段进行系统化研究。

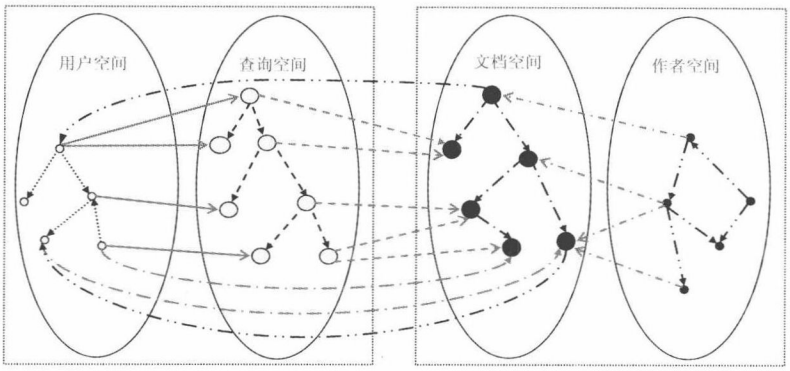


图 1 -2 主题聚类中的空间关系

本书拟以文本聚类为研究对象(将查询式看成一种特殊的文本),进行基于主题的文本聚类研究,如图 1 -3 所示,文本聚类包括文本特征提取、聚类、聚类结果描述 3 个步骤。

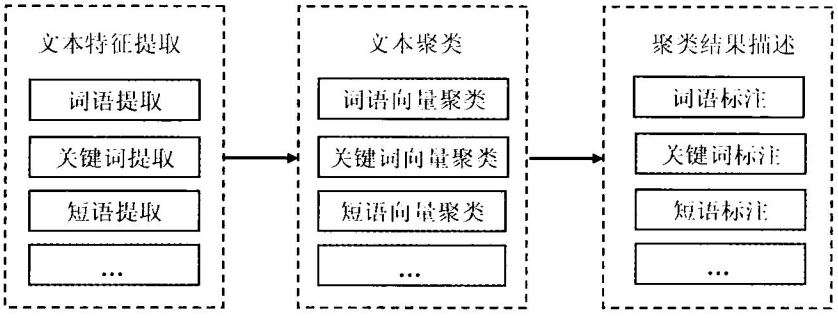


图 1 -3 文本聚类流程图

对文本特征的提取主要可以分为词语提取、关键词提取和短语提取;在文本聚类过程中,分别对应基于词语向量、基于关键词向量、基于短语向量等不同颗粒度文本表示方法的文本聚类;在最后的聚类描述中,分别对应基于词语、基于关键词、基于短语等不同颗粒度的聚类描述。关键词或短语具有结构稳定、语义完整等特点,能从一定程度上克服向量空间模型和概率

独立假设的缺点，更适合作为文本表示的特征，提高文本聚类的效果。[9]需要指出的是，虽然出现了大量的文本聚类算法，但目前对文本聚类结果描述的研究才刚开始。通常，基于词语的文本聚类，聚类过程简单，但聚类结果难以描述，而基于关键词或短语的文本聚类，聚类描述简单，但文本聚类质量较差。[20]目前的文本聚类结果描述大多数是利用简单的统计方法得到最重要的词语，将其直接作为聚类描述。这种方法生成的聚类描述通常可读性不强，不易为用户理解和接受。

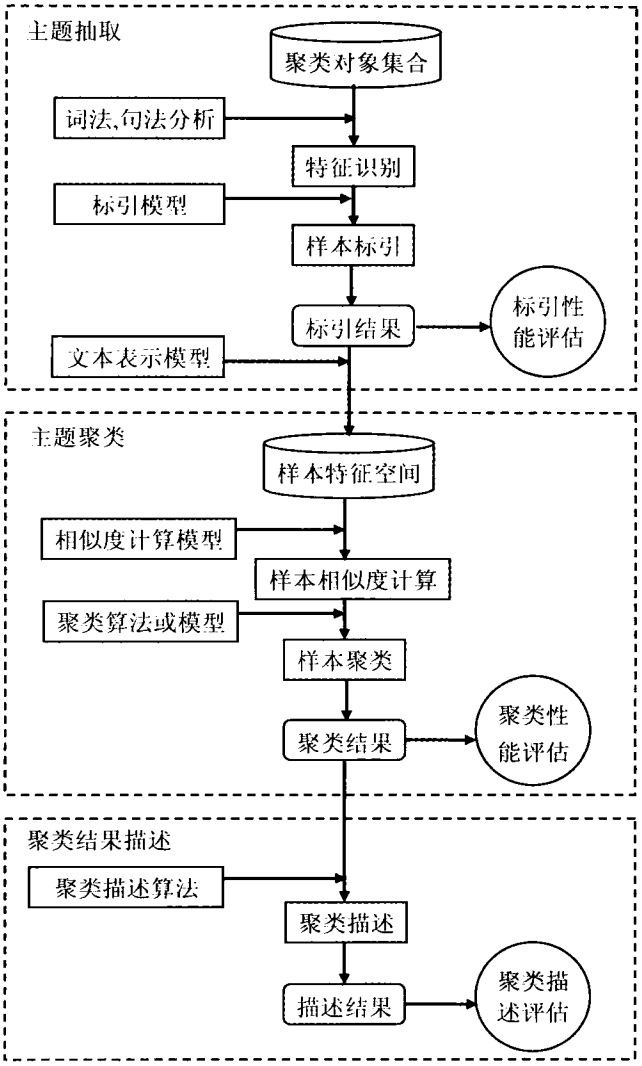


图 1－4　主题聚类的一般流程图

文本聚类是当前研究热点之一。提高文本聚类质量与实用化程度、提高聚类结果描述的可理解性等是文本聚类迫切需要解决的问题。本书从主题角度出发，提出主题聚类方法，并进行主题聚类理论、技术及应用研究。如图 1－4 所示，主题聚类一般包括主题抽取（Topic

Extraction)与聚类(Topic Clustering)以及聚类结果描述(Cluster Description)3个部分,主题聚类的过程一般如下:

① 对聚类样本进行词法、句法分析获得对象特征;

② 根据标引模型对聚类样本进行自动标引,并进行标引性能的评估;

③ 依据文本表示模型,将样本标引结果映射到样本特征空间;

④ 利用相似度计算模型计算聚类样本间的相似度;

⑤ 使用聚类算法或模型对聚类样本进行聚类,并进行聚类性能的评估;

⑥ 通过聚类描述算法获得聚类结果的概念描述,并对描述性能进行评估。

主题聚类过程中存在的5个方面的问题,即如何增强主题提取评估的可靠性并降低主题提取评估成本?如何提高主题提取的实用性?如何提高聚类对象相似度计算的可靠性?如何提高基于样本加权的文本聚类方法的实用性?如何增强文本聚类结果的可读性?本书就这5个问题进行深入研究,创新性结果概括如下:

1. 主题提取评价问题与解决方法

针对常规自动标引评价方法存在评价结果不能完全反映真实标引结果以及评价成本高的情况,本书提出一种通用的自动标引评价模型,该模型有效利用外部资源对标引结果进行有效的评价,增加评价的可靠性并降低评价的成本。

2. 主题提取的实用化问题与解决方法

针对传统的主题提取(包括抽词标引与赋词标引)不能有效利用多个特征及不能完全实用化的问题,本书融合可利用的多个特征,并考虑到标引可以转换为序列标注问题,利用条件随机场模型[21]进行关键词的自动提取研究;融合多个标引模型的标引结果进行投票学习,提出基于集成学习策略的自动标引方法。文本还提出基于Citation-KNN的自动赋词标引算法,提高赋词标引的实用化程度。

3. 聚类对象相似度计算方法存在的问题与解决方法

传统的聚类对象的相似度计算方法往往只考虑对象的一种类型的特征,比如词形上的相似特征,而忽视聚类对象在语义、语境等多个角度上的相似特征,导致它们的相似度计算结果并不可靠。针对该问题,本书提出基于多层特征与多语境的聚类对象相似度计算模型。针对计算字符串相似度传统方法的不足之处,本书提出以相似元作为字符串的基本处理单元,综合考虑相似元的字面、语义及统计关联等多层特征的字符串相似度计算方法。针对传统查询式相似度计算方法的不足,本书利用语料库、释义词典、用户搜索日志作为查询式的不同的语境,进行基于多语境的查询式相似度计算方法,并将该算法用于查询词的相关词的自动获取应用中。

4. 样本加权聚类算法的实用性问题与解决方法

作为一种最近才引起人们注意的算法,样本加权聚类算法[22]还存在一些需要解决的问题,例如聚类对象之间的结构信息对样本加权聚类是否有帮助,如何将结构信息自动转换为样本或对象的权重?针对该问题,本书以学术论文为聚类对象,以K-Means、Fuzzy C-Means算法为聚类算法基础,利用论文之间的引用关系计算每篇论文的PageRank值,并将其作为权重,提出一种基于样本加权的新的文本聚类算法。实验结果表明,基于论文PageRank值加权的聚类算法能改善文本聚类效果。本书利用该算法进行两个方面的应用,即基于主题聚类的主题数字图书馆的设计与实现,基于主题聚类的学科热点的检测。

5. 文本聚类结果可读性问题与解决方法

传统的文本聚类算法只对文本进行聚类，而缺乏对聚类结果进行有效描述的算法，提供给用户的聚类结果可读性与可理解性较弱。针对文本聚类结果可读性较弱问题，本书提出一种增强聚类结果的可理解性与可读性的算法，即基于支持向量机等机器学习方法的文本聚类结果描述算法。为了进一步提高类簇描述词的质量，本书提出一种基于 DCF-DCL 组合策略的文本聚类结果描述算法。实验结果表明这两个算法所取得的效果要优于常规的聚类结果描述方法。本书综合主题提取、文本聚类、聚类描述等方法用于搜索结果聚类这一应用当中。

1.4 本书内容安排

如图 1－5 所示，根据主题聚类中所面临的问题，本书进行 5 个方面的研究工作，它们分别对应于本书的第 3 至第 7 章。

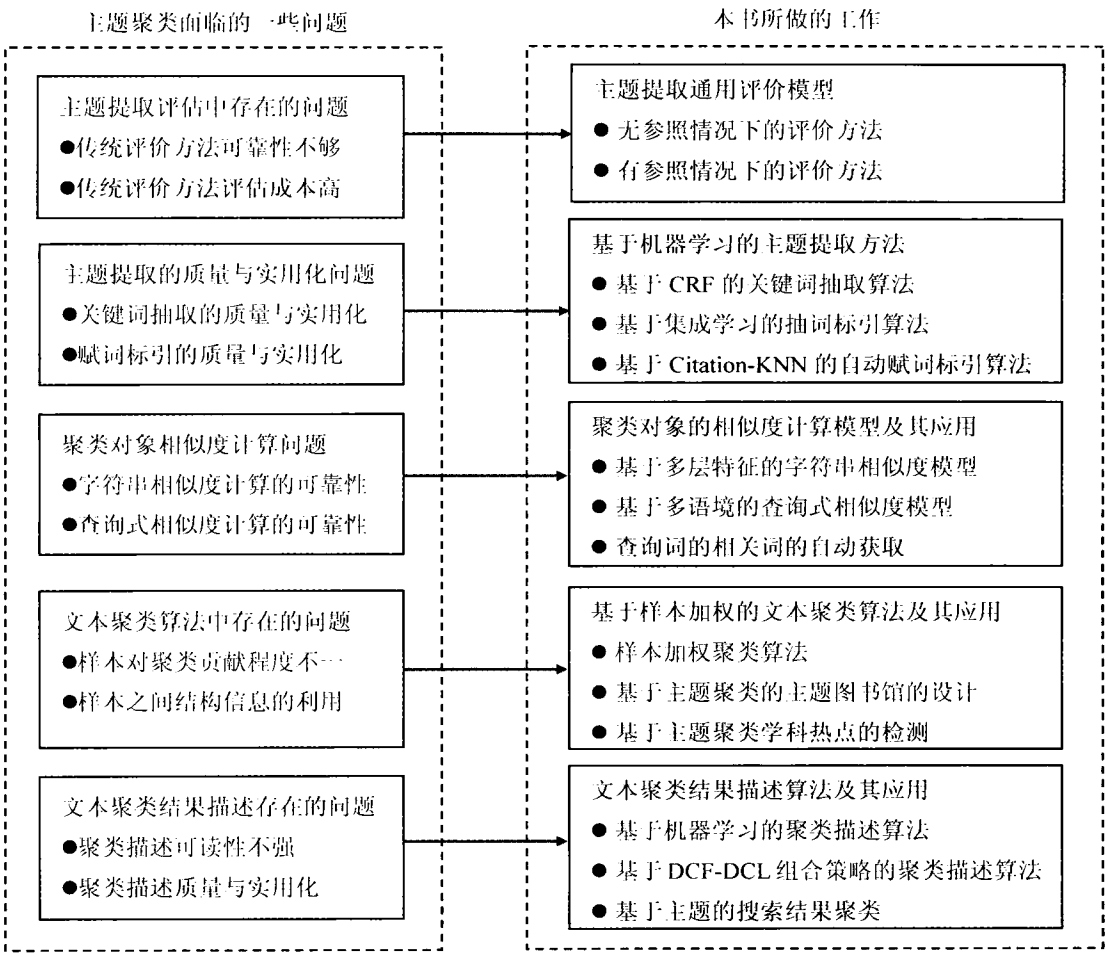


图 1－5　主题聚类面临的问题与本书所做的工作

第 1 章阐述主题聚类的研究背景、研究意义，提出了主题聚类中存在的一些问题，并简要介绍本书所提出的解决方法。

第 2 章对主题聚类中涉及的关键问题进行概述，主要对主题提取、文本聚类、查询式聚类、查询结果聚类的研究现状进行概述。

第 3 章研究主题聚类中存在的第一个问题，即主题提取评价问题，并给出解决方法。本章提出自动标引的通用评价模型，给出有参照与无参照情况下的自动标引结果的评价方法，并给出评价模型的应用和性能分析结果。

第 4 章研究主题聚类中存在的第二个问题，即主题提取的实用化问题，并给出解决方法。本章提出基于机器学习的主题提取方法，详细描述基于条件随机场模型的自动标引算法、基于集成学习的自动标引算法以及基于 Citation-KNN 的自动赋词标引算法，并对这 3 个算法的实验结果进行评估。

第 5 章研究主题聚类中存在的第三个问题，即聚类对象相似度计算方法存在的问题，并给出解决方法。本章提出基于多层特征的字符串相似度计算模型与基于多语境的查询式相似度计算模型，并用于用户查询词的相关词自动获取应用中。

第 6 章研究主题聚类中存在的第四个问题，即样本加权聚类算法的实用性问题，并给出解决方法。本章提出基于样本加权的文本聚类算法，对该算法的实验结果进行评估，并将该算法用于主题数字图书馆与学科热点检测应用中。

第 7 章研究主题聚类中存在的第五个问题，即文本聚类结果可读性问题，并给出解决方法。本章提出基于机器学习、DCF-DCL 组合策略的聚类描述算法，并综合利用主题提取、文本聚类、聚类描述等方法应用于搜索结果聚类。

第 8 章在总结分析全文研究成果的基础上，提出需要进一步研究的问题。

附录给出本书研究过程中所使用的数据与实验样本的例子以及应用结果样例。

参考文献

[1] 侯汉清，马张华. 主题法导论. 北京：北京大学出版社，1991：1.

[2] 马张华. 信息组织. 北京：清华大学出版社，2003：185.

[3] 侯汉清，马张华. 主题法导论. 北京：北京大学出版社，1991：5—8.

[4] 张燕飞. 信息组织的主题语言. 武汉：武汉大学出版社，2005：3—4.

[5] Lahtinen T. Automatic Indexing: an Approach Using an Index Term Corpus and Combining Linguistic and Statistical Methods. Academic Dissertation, University of Helsinki, Finland, 2000: 83—97.

[6] Medelyna O. Automatic Keyphrase Indexing with a Domain - Specific Thesaurus. Master Thesis, University of Freiburg, Germany, 2005: 5—6.

[7] Moens M F. Automatic Indexing and Abstracting of Document Texts. Boston/Dordrecht/London: Kluwer Academic Publishers, 2000: 1—22.

[8] Sebastiani F. Machine Learning in Automatic Text Categorization. ACM Computer Survey, 2002, 34(1): 1—47.

[9] Leouski A V, Croft W B. An Evaluation of Techniques for Clustering Search Results. Technical Report IR - 76, Department of Computer Science, University of Massachusetts, Amherst, 1996: 1—19.

[10] 李航. 文本数据挖掘. 见：王珏，周志华，周傲英主编. 机器学习及其应用. 北京：清华大学出版社，2006：225.