医学文献集合的主题抽取和主题聚类实践*

□ 殷蜀梅 / 北京大学医学图书馆 北京 100083
□ 张智雄 / 中国科学院国家科学图书馆 北京100190

摘要：文献中的重要关键词能够反映其核心主题，因此对文献主题的发现和抽取问题就转化为对文献中的重要关键词集合的抽取。文章在调研了国外在主题抽取和聚类方面采用的技术方法的基础上，提出了在医学学科领域从文本信息资源中抽取主题并进行主题领域判断的技术方案，并详细阐述了其中的主题聚类的技术环节。为了验证该技术方案的有效性，文章以骨关节炎领域为例，对文中提出的技术方案进行实践验证。验证的结果表明文章提出的技术方案有着实际的有效性。该文为2008年第9期本期话题“知识抽取”的文章之一。

关键词：知识抽取，主题抽取，BM25F，MMTx，文本挖掘，医学数据挖掘，数字图书馆
DOI:10.3772/j.issn.1673-2286.2008.09.005

1 引言

一篇学术文献可以包含多个关键词，但只有其中一部分重要的关键词能够表达文献的主要内容。这组重要关键词对于文本有着重要的强文本表示功能。所谓强文本表示功能是指“在文本表示时，能将文本的内容特征（例如领域类别、主题思想、中心意义等）鲜明地表示出来”[1]。这部分重要的关键词也就是这篇文献的主题。通过对多篇文献主题的分析，人们可以从中发现隐含的热点主题领域。一个主题领域往往包含着丰富的内涵，以单个主题词来反映一个主题领域有其局限性，以多个相互关联的主题来描述主题领域比以单个主题词更为直观和全面。知识抽取就是从海量的文献集合中发现隐含的知识，因而主题的抽取和主题领域的聚类对于知识抽取具有重要的意义。在利用计算机进行主动地知识抽取的过程中，如何从文献中抽取文献主题以及如何从众多的文献中发现研究的主题领域是亟需解决的两个重要问题。

本文针对主题抽取和聚类这两个关键问题，在调研目前国外采用的技术方法基础上，提出在医学领域从文本信息资源中抽取主题并进行主题领域判断的技术方案，其中具体的主题抽取技术细节已另外撰文详细说明，本文在提出主体抽取和聚类的技术框架后，以主题聚类的技术细节为重点进行详细阐述，并以医学中骨关节炎学科领域为例进行了实证研究。

2 当前主题抽取和聚类的技术方法

主题抽取和聚类是当前文本挖掘研究的热点之一。各个相关项目对于如何抽取主题并形成主题领域都提出了各自的理解和方法。具体的技术方法有：

（1）从高频被引论文中抽取高频词来代表主题领域

ISI[2]以5年的高频被引论文和核心文献为基础，利用论文共引理论，以双引聚类方法聚类同时被一篇或几篇论文引用的文献类群，然后根据划分的专业进行专业聚类，产生各学科的专题文献束，从这些文献束的文献题名中统计出出现频次较高的、能够反映科学前沿动态的一系列词簇，每个词簇对应着若干具有共引关系的核心文献。

但是这种方法在主题抽取和主题领域判断方面

* 本文受国家社会科学基金项目“从数字信息资源中实现知识抽取的理论和方法研究”（05BTQ006）和国家“十一五”科技支撑计划课题“网络科技信息监测与评价”（2006BAH03B05）的资金资助。

有一定的缺陷：在论文的引用活动中存在着一些非正常的引用，这些引用行为不能真实地反映科学发展、交流的过程；对论文的引用行为是一个已经发生的行为，高频被引论文只能反映已经被大家所重视的主题，只能反映历史而不能反映未来，对于潜在的研究趋势即正处于上升状态的研究趋势不能及时地反映；采用引文分析方法界定主题领域不可避免会有一定的时滞性，由于文献从创作到发表再到被数据库收录需要一定的时间，再到被其他学者引用所需要的时间也更长。

（2）基于语义局部性思想来界定主题领域

美国Lehigh大学开发的HDDI[3]算法中抽取文本信息中的名词短语作为概念词，并提出以语义局部性现象来聚类抽取的概念词形成主题领域。

首先从文本中抽取复杂名词短语作为概念词，然后根据概念词间的共现关系形成概念词关联关系的有向图，并修剪掉重复的关联关系，从有向图中发现紧密关联的区域，舍弃这些区域中概念词数量小于等于2的区域，这样得到的其他区域即视为主题领域。

（3）以词频变化率处于突发状态的主题词作为主题领域

Citespace采用Kleinberg[4]的算法认为“话题报道的数量不是随着时间平滑的增长，而是在不同数量状态之间跳跃。它认为如果一个主题处在不断增长的状态，那么该主题正受着研究者的关注，在聚集越来越多的力量，该主题即使没有达到高频词的要求也可以认为是未来的发展方向”。该算法通过利用“概率机对不同时间段上主题出现的频次进行建模，概率机的状态确定了某时间点上主题出现频次的期望值，而概率机的状态改变由概率模型控制。词突发时，概率机处于高频状态”[5]。在人为确定“状态的个数、状态差异的大小以及状态改变的成本等参数的基础上，利用Viterbi动态建模法对状态改变的概率模型求最优解，即可以得出概率机状态变化的最优时序序列”[5]。

从上面的分析我们可以看出，由单个主题词来代表研究方向有其局限性，目前新兴的方法多是通过将关联紧密的关键词通过某种方法聚合在一起，通过研究多组紧密关联的关键词在信息资源中表现出的属性特征来判断主题领域。

3 主题抽取和聚类形成主题领域的整体思路和框架

在研究了医学领域信息资源的特点以及国外相关主题抽取和聚类的技术方法之后，我们设计了主题抽取和聚类形成主题领域并对主题领域进行评价来分析新兴研究趋势的整体思路和框架，如图1所示：

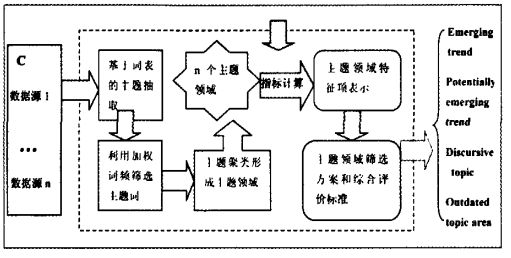

图1 主题抽取和聚类形成主题领域并进行评价的整体思路和框架图

3.1 各个模块功能评述

主题词抽取是从来源文档集合中抽取出文章重点讨论的专业主题词。该部分依据对文章中文本信息的语义单元切分，将文献中能够反映文献主要内容的主题词抽取出来，根据医学专业词表医学一体化语言系统（Unified Medical Language System，UMLS）将文献中的自由文本词映射到UMLS中的专业主题词，将同义词用专业且唯一的主题词来表示。完成对主题词的抽取后，我们需要进一步对抽取出来的主题词或词组进行加权词频统计，从抽取出的众多主题词中挑选能够反映文献主要讨论内容的主题词集合。为了消除泛义词对趋势分析的影响，首先应该对抽取出的主题词根据其对文章实际内容贡献程度大小来筛选主题词，将能够真正表达文献主要讨论内容的主题词挑选出来。相关的详细技术细节参见“一种对医学文本中重要关键词抽取和筛选的技术方法”一文。

主题词聚类是通过主题词在文献集合中的共现强度来进行聚类，形成n个重点主题领域。词共现是指如果两个主题词在同一篇文献中同时出现，则认为这两个主题词间存在着某种关联；如果同时包含这两个主题词的文献越多则说明这种关联程度越紧密。我们知道单个主题词不能全面反映一个研究方向，在从文

本信息资源中提取出多个关键词的基础上，需要将关联紧密的关键词聚合成类形成主题领域，才能全面地代表一个具体的研究方向。因此需要通过基于词共现的聚类方法使每个主题领域内部的主题词互相紧密关联，而主题领域之间则相对独立。

主题领域评价是在充分了解新兴研究趋势所应具备的各项特征的基础上，结合现有信息资源所能够提供的信息来综合设计评价指标。借助得到的主题领域及其在评价指标体系上的表现特征，我们进一步综合这些研究结果进行判断。一个主题领域的研究情况是随着时间动态变化的，要进行研究趋势的判断就需要对评价指标的时间序列进行分析，综合评价指标体系中的各项指标，对其进行抽象，归纳为主题领域的表现特征。依据主题领域表现特征在时间序列上的变化，通过分析时间序列进行综合判断。根据每个主题领域在每个指标上的不同表现情况，将候选主题领域划分为新兴研究趋势（Emerging trend）、潜在新兴研究趋势（Potentially emerging trend）、分散的主题领域（Discursive topic）和不再流行的主题领域（Outdated topic area）。

限于篇幅，本文重点阐述主题聚类的技术方案。

3.2 主题词聚类技术方法选择

聚类就是将数据对象分成类或簇的过程。当前的聚类分析的方法有多种，主要有层次方法（hierarchical method）、划分方法（Partitioning method）、基于密度的方法（density-based method）、基于网格的方法（grid-based method）、基于模型的方法（model-based method）等等[6]。

结合本文根据主题词间共现程度大小进行聚类的要求，聚类方法中的划分方法较为符合本文的要求，其中心思想是：随机选择k个对象，每个对象初始地代表一个类的平均值或中心，对剩余每个对象，根据其到类中心的距离，被划分到最近的类；然后重新计算每个类的平均值。不断重复这个过程，直到所有的样本都不能再分配为止。划分方法可以根据对象间关联关系远近来聚类，并通过逐步修正每个聚类的中心来达到使得类的内部关联关系紧密而类之间相对独立的目的。本文采用划分方法中较为典型的一种算法——K-means聚类算法。

3.3 主题词聚类的技术方案

采用k-means聚类算法，首先要明确关键词间的关联关系远近的计算方法，以关联关系的紧密程度来代表每个关键词间的距离。为了尽量从语义角度来衡量关键词间的关联程度，本文拟采用以关键词共现程度来衡量两个主题词之间的关联程度。

关键词共现程度衡量算法采用Salton指数。“S指数可以使两个本来关系就密切的关键词显现得更密切，使关系疏远的关键词显现得更为疏远”[7]。即将共词矩阵中每个数字除以与之相关的两个主题词出现频次乘积的开方。计算公式：$S=n_{ij}/\sqrt{n_i*n_j}$，其中，n_i和n_j分别表示主题词i和j的频次，n_{ij}表示主题词i和j共现的频次。

本文对主题词i和主题词j间的关联关系距离大小distance(i,j)计算公式为：

$$distance(i,j)=\begin{cases}n_{ij}/\sqrt{n_i*n_j} & 当n_{ij}>0\\ 10000 & 当n_{ij}=0\end{cases}$$ 公式3-1

根据k-means算法和主题词间关联距离远近的计算方法，笔者拟定如下聚类运算步骤：

步骤1：在主题词集合$Theme_{main}$中选定K个聚类中心，初始每个类的聚类中心从主题词集合中随机抽取k个主题词作为k-means聚类的初始中心；

步骤2：将全部主题词按与类中心的距离最小原则分到K个类中，该距离的衡量按照上述的公式3-1计算；

步骤3：重新计算每个类的中心词（以与其他同类主题词之间的距离之和为最小的主题词为其每个类的中心）；

步骤4：以上述第3步求出的新的聚类中心为中心；

步骤5：如果新的聚类中心=原来的聚类中心，则结束，否则转2。

根据以上的步骤，笔者设计相应聚类流程图如图2所示。

4 具体的试验和评价

本文以Medline和SCI数据库为数据来源，从中获取医学文献的文摘以及被引用次数。在研究领域的选择方面，笔者咨询了北京大学医学图书馆的学

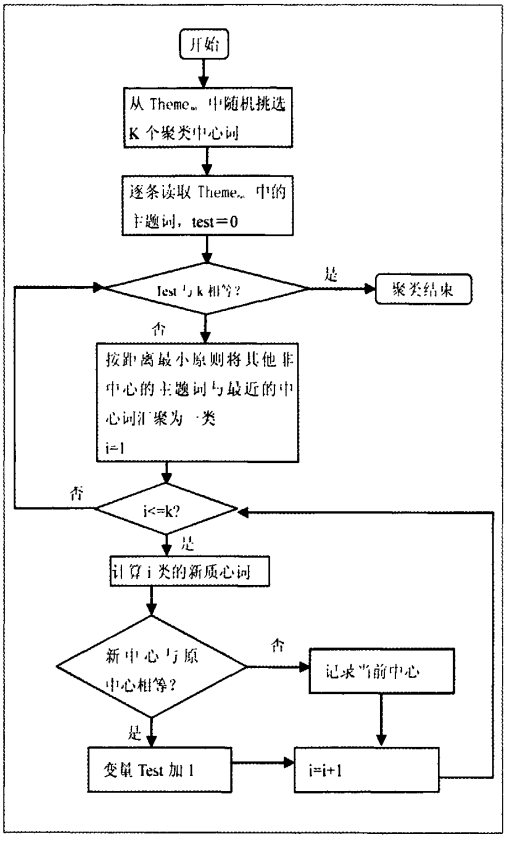

图2 利用k-means方法聚类主题词的流程图

科馆员，最终将骨关节炎疾病作为进行试验的领域。通过题名匹配的方式进行数据清洗后，SCI与Medline在骨关节炎相关的文献集合中能完全匹配的记录有6530条。本文依据该6530条的文献集合进行骨关节炎研究领域的主题领域聚类。

依据以上设计的技术方案，通过具体试验，我们一共聚集了20个主题领域，表1显示每个主题领域的主要内容和重要主题词。

5 结语

我们对这20个主题领域进行分析和评价，根据其在不同评价指标上的表现特点，从中发现了正吸引研究者注意的主题领域：骨关节炎的生物学标记研究、骨关节炎疼痛的缓解、膝关节置换术等。根据相关专家的意见和相关文献的收集，我们的判断结果与文献综述的观点和专家的意见在很大程度上是一致的。

通过上述的研究，我们提出了从医学文本中自动进行主题抽取和聚类的技术方法，并通过实际的验证说明其是切实可行的。这对于挖掘医学文本有着重要的意义，为文本挖掘中主题的抽取和聚类问题提供一种切实可行的技术方法。

表1 20个主题领域标号与中文译名对照表

序号	主题领域内容（中文解释）	重要主题词（部分）
1	氨基葡萄糖、硫酸软骨素在治疗骨关节炎中的机制	GLUCOSAMINE(氨基葡萄糖)，CHONDROITIN(软骨素)，CHONDROITIN SULFATE(硫酸软骨素)，CHONDROITIN SULFATES(硫酸软骨素)，
2	基质金属蛋白酶的合成机制	SYNTHESIS(合成)，MMP1(基质金属蛋白酶1)，MMP3(基质金属蛋白酶3)，MMP-13(基质金属蛋白酶13)
3	骨关节炎的生物学标记研究（生物示踪技术在骨关节炎的应用）	CARTILAGE(软骨)，COLLAGEN(II型胶原)，PROTEOGLYCANS(蛋白多糖)，TISSUE(组织工程)
4	膝关节炎病程进展导致膝关节疼痛、弓形腿、外翻足	OSTEOARTHRITIS，KNEE(膝关节炎)，VARUS(内翻)，DISEASE PROGRESSION(病程)
5	骨关节炎疼痛的缓解	PAIN(疼痛)，ANALGESICS(止痛剂)，PAIN RELIEF(疼痛缓解)
6	老年人与髋关节炎的关系	OSTEOARTHRITIS，HIP(髋关节炎)，ELDERLY(老年人)
7	膝关节置换术	TOTAL KNEE ARTHROPLASTY(全膝关节置换术)，
8	性别与年龄在骨关节炎发病的区别	OA(骨关节炎)，SEX(性别)
9	老年人骨关节炎后期通过锻炼恢复治疗	EXERCISE(锻炼)，REHABILITATION(复原)，
10	前十字韧带损伤与骨关节炎的关系动物实验模型	DOGS(狗)，ANTERIOR CRUCIATE LIGAMENT(前十字韧带)
11	肿瘤坏死因子-α与骨关节炎的关系	TNFALPHA(肿瘤坏死因子α)，
12	通过外科手术（如关节造型术、关节固定术、骨切开术、关节置换术等）治疗骨关节炎	SURGERY(外科)，ARTHROSCOPY(关节镜检查)，JOINT REPLACEMENT(关节置换)

13	肥胖是骨关节炎的高危因素	OBESITY(肥胖), HYPERTENSION(高血压), BODY MASS INDEX(BMI),
14	透明质酸治疗骨关节炎的机制	HYALURONIC ACID(透明质酸),
15	骨关节炎中关节软骨损伤的动物实验模型	ARTICULAR CARTILAGE(关节软骨), DAMAGE(损伤), HORSES(马),
16	骨密度与骨关节炎的关系	BONE(骨), BONE MINERAL DENSITY(骨矿物质密度),
17	风湿性关节炎与骨关节炎的关系	RHEUMATOID ARTHRITIS(风湿性关节炎),
18	骨关节炎基因疗法中通过抑制炎性细胞因子、基质金属蛋白酶的表达以及促进生长因子的表达的研究	EXPRESSION(表达), CYTOKINES(细胞因子), IL1-BETA(白细胞介素), MATRIX METALLOPROTEINASES(基质金属蛋白)
19	传统非甾体类抗炎药治疗骨关节炎	CYCLOOXYGENASE INHIBITORS(环加氧酶抑制剂), NSAIDS
20	手关节炎	HAND(手), HAND OSTEOARTHRITIS(手关节炎)

参考文献

[1] 刘华. 基于文本分类中特征提取的领域词语聚类[J]. 语言文字应用, 2007(1):139-144.

[2] Essential Science Indicators[EB/OL]. [2007-08-01]. http://www.esi-topics.com/RFmethodology.html.

[3] POTTENGER W M, KIN Yong-Bin, MELING D D. HDDITM: Hierarchical Distributed Dynamic Indexing[EB/OL].[2007-08-01]. http://www.cse.lehigh.edu/~billp/pubs/HDDIFinalChapter.pdf.

[4] KLEINBERG J. Bursty and hierarchical structure in streams[EB/OL]. [2007-08-01]. http://www.cs.cornell.edu/home/kleinber/bhs.pdf.

[5] 魏晓俊. 基于科技文献中词语的科技发展监测方法研究[J]. 情报杂志, 2007(3):34-39.

[6] 数据挖掘中聚类分析的技术方法[EB/OL]. [2007-08-01]. http://bidwhome.itpub.net/post/20871/155927.

[7] 梁立明, 武夷山. 科学计量学: 理论探索与案例研究[M]. 北京: 科学出版社,2006.S

作者简介

殷蔚梅 (1977-), 女, 北京大学医学图书馆, 馆员, 发文7篇。通讯地址: 北京市海淀区学院路38号北京大学医学图书馆 100083

张智雄 (1971-), 男, 中国科学院国家科学图书馆研究馆员、博士生导师, 发文60余篇。通讯地址: 北京市海淀区中关村北四环西路33号, 中国科学院国家科学图书馆 100190

A Method for Topic Extraction and Clustering Based on Medical Literature

Yin Shumei / Peking University Health Science Library, Beijing, 100083
Zhang Zhixiong / National Science Library, Chinese Academy of Sciences, Beijing, 100190

Abstract: Important keywords in academic papers reflect topics of the literature. Therefore, the extraction of topics turns to be the extraction of keyword groups. This paper first investigates techniques for topic extraction and clustering used by overseas, then the researchers propose a technical scheme for extracting topics in text information resources in the medical field and for topic area identification. A detailed explanation of the techniques for topic clustering is given. To verify the validity of the method, this paper applies the scheme to the field of osteoarthritis research. The result proves the validity of the proposed method.

Keywords: Knowledge extraction, Topic extraction, BM25F, MMTx, Text mining, Medical data mining, Digital library

(收稿日期: 2008-07-13; 责任编辑: 贾延霞)

(上接12页)

Review of the Technologies and Methods for Extracting Content Objects from Unstructured Text

Zhang Zhixiong, Wu Zhenxin / National Science Library, Chinese Academy of Sciences, Beijing, 100190
Zhao Qi, Hong Na / National Science Library, Chinese Academy of Sciences, Beijing, 100190;Graduate School of the Chinese Academy of Sciences, Beijing, 100049
Xu Jian / National Science Library, Chinese Academy of Sciences, Beijing, 100190; Graduate School of the Chinese Academy of Sciences, Beijing, 100049;Department of Information Management, Sun Yat-Sen Univ., Guangzhou, 510275
Liu Jianhua / National Science Library, Chinese Academy of Sciences, Beijing, 100190; Graduate School of the Chinese Academy of Sciences, Beijing, 100049

Abstract: In recent years, knowledge extraction plays a very important role when dealing with unstructured text. In this paper, based on the analysis of current relevant literature, systems and projects, it proposes the classification of the current knowledge extraction objects and reviews the relevant technologies and methods. The major themes include web object identification and integration, terminology extraction, topic discovery, conceptual hierarchy relation extraction, non-conceptual hierarchy relation extraction, fact extraction and opinion extraction. This paper also analyzes trends of knowledge extraction in the future.

Keywords: Knowledge extraction, Object identification, Terminology extraction, Topic discovery, Relation extraction, Fact extract, Opinion extraction, Digital library

(收稿日期: 2008-07-13; 责任编辑: 贾延霞)