文本挖掘技术在生物医学文献管理中的应用

楼婷渊 孟志青 胡强
(浙江工业大学经贸管理学院 浙江杭州 310023)

摘要：生物医学文献以非结构化的文本形式存在，文本挖掘能够从海量的生物医学文献中发现有趣的知识和模式，可以提高对生物医学文献的管理和建设效率。本文针对生物医学领域，阐述了文本挖掘的具体过程，论述了生物医学文本挖掘现有的研究方法，详细讨论了生物医学文献的分类和关系抽取，最后对文本挖掘在生物医学领域的应用前景做了展望。

关键词：文本挖掘 生物医学文献 文本分类 关系抽取

一、引言

信息爆炸时代，各行业每时每刻都在产生和积累大量的以各种形式保存的信息，这些信息以指数级的速度不断积累和增长，如何快速准确地从这些纷乱的数据中提取出有价值的信息是急待解决的问题。文本挖掘是指从大量文本数据中抽取事先未知的、可理解的、最终可用的知识的过程，同时运用这些知识更好地组织信息以便将来参考[1]。如今文本挖掘已经成为国际上非常活跃的一个研究领域。

随着生物医学领域的快速发展，生物医学文献呈指数级增长，成为一座巨大的知识宝库。然而面对如此大规模的、快速增长的科学文献数据，即便是该领域内的专家也无法依赖手工方式从中获取感兴趣的信息。由于生物医学文献绝大多数都是以非结构化的形式存在于文本文件中，因此采用文本挖掘技术对生物医学文献数据进行管理是非常有必要的。

二、文本挖掘过程

文本挖掘通常包括文本数据预处理、特征信息提取和数据挖掘三个步骤。文本挖掘过程如图1所示：

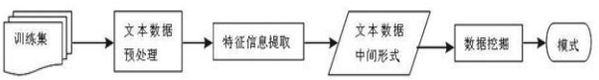

图1 文本挖掘过程

文本数据预处理的质量会直接影响到最终的结果，英文文本数据预处理包括无用词过滤和词干化处理。文本特征信息提取是将非结构化或半结构化的文本数据转化为挖掘工具可以处理的中间形式的过程，特征提取首先要识别文本中包含重要信息的特征项。本文采用数学模型来表示这些特征项，常用的特征表示模型有布尔模型、向量空间模型和概率模型，通过特征表示得到的向量维数较高，特征抽取的基本思想是利用映射的方法将高维特征映射到低维空间中，特征抽取一般是构造一个评价函数，然后对每个特征向量进行评估，删除评估分数较低的特征向量。经过特征信息提取之后，文本数据以结构化形式存储在数据库中，因此计算机就可以对文本数据的特征信息进行分类、聚类、关联分析和趋势分析等数据挖掘处理。

三、文本挖掘技术在生物医学文献管理中的应用

将文本挖掘技术应用到生物医学领域中，通过挖掘文本数据发现生物医学的规律，能够提高生物医学文献管理的效率。

(一)生物医学文献分类

对生物医学文献进行合理分类可以对文献的组织和搜索带来极大的便利，也为进一步的数据处理打下基础。文本分类是指将文本数据映射到预先定义好的类别中，我国常用的分类方法有基于距离的方法、决策树分类法、贝叶斯分类法等。生物医学文献语料库是对生物医学文献分类的基础，目前国际上可以公开获取的生物医学语料库有：GENIA语料库、Yapex语料库、PDG语料库等。另外由于生物医学文献中的专用术语较多，有些术语在文献中出现次数不多但非常重要，具有很强的分类特征，因此如何在已有的分类方法的基础上设计出符合这一特点的算法来提高生物医学文献分类的准确率和效率是亟待解决的问题。

(二)生物医学文献关系抽取

生物医学文献关系抽取的目的是从文献信息中找出生物实体之间的关系，例如基因与某种疾病之间的关系。由于生物医学文献中同一概念有多种不同的表示方法，同时文献中也可能出现很多语料库中不存在的新概念，因此生物医学文献关系抽取的难度较大，国际上常用的关系抽取方法有共现方法、关键词方法、机器学习方法和自然语言处理方法[2]。这些方法在生物医学文献关系抽取中都存在一些不足之处，有学者提出利用向量空间模型来识别文献中生物实体间的关系，在现有方法的基础上进行开发或多种方法融合运用以期获得更准确的关系抽取结果。

本文主要介绍了生物医学文献的分类和关系抽取，当前生物医学文本挖掘的研究热点主要集中在文献分类、信息检索、自动摘要、生物医学领域实体识别、文献信息关系抽取等方面。通过文本分类可以缩小搜索范围，为后续的数据处理做准备；通过信息检索可以帮助用户在海量的文本信息中快速找到有价值的信息；通过自动摘要技术计算机可以自动地从原始生物医学文献中提取出主要内容，使研究者不用花费较多时间就可以从海量的生物医学文献中获得有价值的信息。通过文献信息关系抽取技术可以从生物医学文献中抽取出特定的事实信息，对生物知识网络的建立、生物体关系的预测和新药的研制等均具有重要的意义。

四、总结

文本挖掘是当今国内外学者研究的热点问题，其在生物医学领域的研究具有广阔的应用前景和重要的现实意义。本文概述了在生物医学文献中文本挖掘的具体过程，重点论述了文本挖掘在生物医学文献的分类和关系抽取中的应用和研究状况。文本挖掘技术在生物医学文献管理中的应用在近年来已取得了一定成果，但在很多方面仍需要更深入地研究和探索，文本挖掘技术的提升将会推动生物医学领域的发展进步。

参考文献：

[1]杨斌,孟志青.一种文本分类数据挖掘的技术[J].湘潭大学自然科学学报,2001,23(4):34-37

[2]王浩畅,赵铁军.生物医学文本挖掘技术的研究与进展[J],中文信息学报,2008,22(3):89-98