

基于文本挖掘的可视化研究

■赵丽丽/山东管理学院

摘要: 基于文本挖掘的可视化是一个跨学科的领域,涉及文本分析、数据挖掘、计算机图形学、人机交互以及认知科学等各方面的理论和技术,它能够将文本中的信息先通过文本挖掘的手段处理成文本数据,再经过可视化转换成图形或者图像,并在计算机的屏幕上显示出来,帮助人们抓住隐藏在海量文本之中的信息的本质、结构和规律。本文首先简要地介绍了文本挖掘与可视化分析的重要性以及巨大的前景和潜力;然后系统地介绍了文本挖掘和可视化技术,并讨论了这两类技术的关系以及如何应用;最后对全文进行了总结并展望了这两项技术进行融合之后的应用以及潜力。

关键词: 文本挖掘 信息可视化

在信息时代,互联网上充斥着极为丰富的各类信息资源,这些信息资源良莠不齐,人们要从如此巨大的资源中获取有效的信息十分困难,效率也很低,因此,文本挖掘系统应运而生,它为人们提供了一种从海量信息中分析获取有效信息的手段。这种技术现在已经被应用在地理信息系统、医学影像成像技术以及产品设计等很多不同的领域,受到国内外学术界、商业界以及政府部门越来越多的重视。

1 前言

近些年来,随着人们对计算机的要求越来越多,计算机科学为满足人们需要,以前所未有的速度在快速发展,不断更新换代,它的功能也越来越强大。但是,在快速发展的过程中,我们也面临着诸多挑战,尤其是因为互联网的飞速发展,文本信息越来越多,信息量已经大到了靠人力无法处理的地步。面对如此庞大的文本信息量,如何处理这么多的信息,并让人们能在大量信息中快速接收有用的信息成为了一项紧急而重要的研究项目。文本挖掘技术这时候出现了,它为人们提供了一种从海量信息中分析获取有效信息的手段。与传统的数据挖掘不同,文本挖掘处理的信息形式主要是文字,由于文字大多是非结构化的,而且语义语法之间有很

多不同,一词多义,多词同义等现象大量存在,因此文本挖掘技术要比数据挖掘技术更加困难。挖掘文本信息就是寻找一些文字的规律性,并且,分析该文字本身与语义、语法之间的联系,从非结构化的文本数据中发现所需要的知识,一般情况下,会采取一些特殊的方法来进行辅助分析,例如,信息提取,文本分类,文本聚类等。而信息可视化能够利用人们对图形图像的超强感知能力,通过可视化数据挖掘,将文本信息变为图像信息,发现隐藏在数据和信息背后的含义,容易被人们接受。

2 文本挖掘

所谓的文本挖掘,就是在概率论与数理统计的基础上,在海量的文本信息中,提取出来未知的,人们所需要的知识。由于要对文本信息进行整合,也就是对词语的词频、词义等等进行统计分析,所以借助一定的数学手段必不可少。

文本挖掘的主要用途是从原来未经处理的海量的文本信息中提取出来未知的知识。但是文本挖掘困难的地方在于它必须处理那些比较模糊而且形式是非结构化的文本数据,所以这就决定了它必须要用到多种技术,涵盖范围包括了信息技术、文本分析、模式识别、数理统计学、数据库技术、深度学习以及数据挖掘等等。文本挖掘目前已经应用到了多个领域。例如,在面对商家发布的大量的垃圾邮件时,我们进行逐一查看十分浪费时间,这时,文字挖掘就发挥出了巨大的作用,通过文字挖掘,将垃圾邮件进行过滤,并把其他邮件进行归类,然后根据邮件内容,进行自动回复,为我们节省了很多时间。另外,在论文查重领域,文字挖掘也发挥了举足轻重的作用,提交的论文经过计算机算法的过滤,将论文数据库中有重复的部分都提取出来,并将重复字数与总字数相比,计算出重复率,并能将重复部分与原文章进行对照,方便作者进行修改。

3 基于文本挖掘的可视化

文本信息可视化主要是指将纷繁复杂的海量的文本信息通过文本挖掘的手段提取出有效的信息,再通过可视化的图形来表示分析计算出的数据,以

达到帮助人们迅速了解有用信息,提高效率的目的。文本内容的可视化重点关注如何快速获取文本内容的重点,先通过文本挖掘的手段,将关键信息提取出来,主要可以分为基于词频的提取和基于词汇分布的提取方式,在可视化进程中的主要方式也可以分为基于词频和基于词汇分布的方式。基于文本内容的可视化既可以应用于单个文本,也适用于多个文本组成的文本集。通过可视化的方式来呈现基本的统计结果,能帮助人们迅速地了解文本的主要内容,这对于下一步的分析具有重要的意义。针对海量的文本信息,文本可视化能够最大程度的对文本信息实现抽象和概括。它呈现的形式不仅仅是丰富多样的图形、图表的组合,更重要的是,发现一篇文档或者一系列文档集合中特定的、潜在的数学结构。因此,我们可以利用计算机的手段来分析处理文本的可视化,这种潜在的数学结构可以帮助我们进一步的分析文本。

4 结语

文本挖掘和信息可视化是两种技术,通过这样两种技术的结合使用,就能够帮助用户充分探索隐藏在海量文本信息背后的含义,从而发现在文本信息中潜在的特点、规律。这两种技术的结合将高度概括并且形象化表示文本信息中的核心内容,方便人们快速的理解和吸收文本中的核心思想。目前,这种技术已经广泛应用到了人们生活中的各行各业,分布在科学研究、医疗卫生、教育教学、地理空间信息传输、产品设计、软件开发、工程制图等等众多领域,有着巨大的发展和应用潜力,极大地提高了人们的工作效率,促进其他学科的蓬勃发展,为社会带来巨大的效益。

参考文献:

- [1]刘春艳. 基于信息可视化的文本挖掘研究领域前沿与演化分析[J]. 图书情报工作 2011(S2): 270 ~ 272, 189.
- [2]王丽坤,王宏,陆玉昌. 文本挖掘及其关键技术与方法[J]. 计算机科学, 2002(12): 12 ~ 19.