论数据挖掘技术的应用

【摘要】

2010 年 9 月我参加了新闻总署的“网络舆情监管信息决策系统”的设计与开发，我担任了系统架构工作，并参与了部分功能代码地编写工作。网络舆情是一个新生事物，网络舆情监管主要指以 BBS 论坛、博客、各种社交网站和虚拟社区等为平台而呈现出来的网民对社会上的人和事的看法监控分析。本项目的设计与开发是通过运用先进的数据挖掘技术，情报技术，内容管理技术，对网络相关舆情进行深入细致地搜集、整理、分析，并对一段时间内的内容进行有效地统计报告，给出对特定关注对象的内容地相关评价，为相关部门决策提供有力地支持作用。我在项目中使用了数据挖掘技术的关联分析，聚类分析，分类分析，预测等方法从舆情知识库中抽取出不同知识库对象，并根据统计分析方法形成相应的决策库系统。我从功能上将整个系统分为“采，编，发”三个重要部分，并选用三层 C/S 结构作为决策系统软件的结构。在系统开发过程中我选用.net 作为系统的基本开发环境，因为它很好地支持 C++等各种开发语言。目前该系统已经交付用户使用，正在为新闻总署净化网络环境提供了可靠的技术支持。

【正文】

2010 年 9 月我所在的单位承接了一个新闻总署的“网络舆情监管信息决策系统”的设计与开发，我在项目中担任了系统架构和部分功能代码地编写工作。网络舆情是近几年的一个新生事物，它是伴随着互联网，Web2.0 技术的发展而逐步发展起来的。网络舆情监管主要指以 BBS 论坛、博客、各种社交网站和虚拟社区等为平台而呈现出来的网民对社会上的人和事的看法监控分析。舆情情报的内容主要是指对网络信息的采集，信息分类，信息统计，敏感信息判定，预警等。在进行舆情分析的过程中需要使用大量的信息规划方法，数据挖掘技术，情报技术和内容管理技术等。本项目的任务是通过对网络相关舆情进行深入细致地搜集，整理，分析，对一段时间内的内容进行有效地统计报告，对特定关注的对象内容给出相关评价分析，为相关部门决策提供强有力地支持作用。

考虑到项目实际的情况，我采用了三层 C/S 结构作为决策系统的系统架构。系统从下往上依次是数据存储层，功能层，表示层。数据存储层的主要任务是负责数据的存储，功能层包含三个部分，一个部分是负责数据处理的智能代理，这个部分包含了数据挖掘的所有处理逻辑，就如同人的大脑一样。另一部分是负责数据分配的应用服务器，这个部分主要是对处理任务进行分配，对客户端程序访问个数进行控制，并起到负载均衡，分布式处理的作用，它如同人的神经，起到连接上下层的作用。还有一部分是负责编辑处理关联规则的规则编辑器。在表示层的任务主要是负责显示，担负着用户和系统之间的交互。整个系统从功能上又可以分为“采，编，发”三个部分。“采”是指从网上采集相关的信息；“编”指负责将采集到的数据经过分类，再经过聚类方式存入相对应的对象库，构成面向方面或者面向领域的数据模型结构；“发”是通过发布的方式，向不同领域的用户提供所关注的对象数据帮助其进行分析决策。下面将通过介绍整个系统设计的工作流程，来具体说明数据挖掘技术在整个系统中应用，并讲述数据挖掘的方法主要有哪些，我在设计架构整个系统过程中是如何进行选择应用的。

数据挖掘的主要任务主要是关联分析，聚类分析，分类分析，预测分析和偏差分析等。整个系统我们通过关联分析建立一个模型，该模型没有采用自学习的方式，而是通过领域构建的过程。假设存在众多地历史文本，在历史文本中经过抽样统计，得到一些关键的词频。由这些词经过规则编辑器可以编辑成为一些词汇的集合。一条规则是由基本的与，或，非，异或逻辑关系通过排列组合而成。几条规则经过排列组合可以构成更为复杂的规则集。首先我们根据词频信息，采集网上相似度较高的文本，这个文本集是粗粒度的原始数据源，通过智能代理聚类学习方式，可以将这些粗粒度的原始数据源进行聚类处理。在聚类的过程中数据彼此相似，不同类中的数据相异。这样可以从宏观上把原始数据进行一次处理，我们称这个过程为“打标签”。 在聚类的过程中是有很多方法可以选择，比如基于 K-means 的方法，还有就是当下最为实用的基于贝叶斯的方法，我们通过选择比较发现基于贝叶斯的方法最终的处理的结果误判率更低，效果更好，所以采用该种方法。经过这次“打标签”处理过后得到的数据就比较有规律性了，根据前面词频关联规则学习过程得到的规则集，再通过智能代理中对当下的二次数据源中所包含的标签信息进行适当的分类学习，分类的过程还是通过计算文本之间的相似度。通过这个过程最终可以生成一个面向领域的知识库。在整个分类过程中我还选用了决策树的方法，决策树是一种预测模型算法，它通过将大量数据有目的的分类，从中找到一些有价值的，潜在的信息。它的主要优点是描述简单，分类速度快，特别适合大规模数据处理。

通过运用多线程技术将最终结果生成一个一个任务，并最终由应用服务器统一调配存入相应的数据表或者数据库中。整个系统选用支持向量机技术表示待处理的文本信息，通过基于构件的技术进行分布式存储。对于数据挖掘方法的选取方面，因为数据挖掘有很多种方法，比如神经网络的方法，遗传算法的方法，粗集方法，统计分析方法，模糊集方法。目前在应用领域中可以说主流是通过基于统计处理的方法和基于概率的处理方法。在我的整个系统架构中，两种方法我都有涉及，比如说在聚类时可以通过基于 K-Means 的处理方法，这种方法就是基于统计的计算方式。这种方法的优点是算法简单，但是处理大规模数据速度效率较差。而基于贝叶斯的方法是基于概率的处理方式，这种方式处理效率高，数学模型简单。所以最终系统选用该算法实现了数据聚类过程。

最后是对整个知识库的发布过程，通过编写的客户端程序我们已经生成的面向领域的决策系统数据库进行数据访问操作。考虑到信息安全等多方面的因素，我们采用了 C/S 的软件结构。如果以后用户提出外部网络的访问需求，还可以在现在的基础之上进行混合结构改造，当然在这个改造的过程当中，对于改变系统架构的工作量是相对要少得很多，因为“发”的过程不涉及到数据的存储和数据的处理过程。 当然设计整个系统的架构过程中我们还考虑到了偏差分析，因为在实际的处理过程当中会出现很多异常的数据，比如说在聚类的过程当中，单次计算过程中会产生一些奇异点数据，对于这些数据我们应该采取丢弃措施。当然对于选用不同的算法计算的结果也有可能是不同的，这样也需要我们进行偏差分析。为了进行偏差分析，我们引入搜索中的查全率和查准率两个概念进行分析的。查全率，它是指检出的相关文献量与检索系统中相关文献总量的比率，它是衡量信息检索系统检出相关文献能力的尺度。查准率，它是指检出的相关文献量与检出文献总量的比率，是衡量信息检索系统检出文献

准确度的尺度。根据计算所得这两个数值对知识库中的数据偏差进行适当的纠正。

目前该系统已经交付用户使用，正在为新闻总署净化网络环境提供了可靠的技术支持。得到了新闻总署领导的一致好评。