

Data Mining Technique of Analysis Software for Patent Map

Chunyu MIAO¹, Lina CHEN²

¹Zhejiang Normal University, Jinhua, China

²Zhejiang Normal University, Jinhua, China

Email: netmcy@zjnu.cn, chenlina@zjnu.edu.cn

Abstract: Patent map is an important tool to collect, arrange and use patents. It plays an important role in patent strategem. It is very difficult to cluster, analyze and predict thousands of information because the quantity of the patent data is a great deal. Data mining is a new technique for information processing and it is becoming increasingly important in usage and extraction for information. We explain the concept and inherent meaning of the data mining in this paper. All kinds of techniques and methods are described. The application that the data mining is used to explore the patent map is further analyzed. The method and flow of the data mining for the patent data are explored.

Keywords: patent analysis; patent map; data mining

专利地图分析软件中的数据挖掘技术

苗春雨¹, 陈丽娜²

¹浙江师范大学, 金华, 中国, 321004

²浙江师范大学, 金华, 中国, 321004

Email: netmcy@zjnu.cn, chenlina@zjnu.edu.cn

摘要: 专利地图作为一种搜集、整理和利用专利信息的重要工具, 其在专利战略中发挥着不可忽视的作用。但专利数据属于海量信息, 想对成百万的信息进行聚类和分析及预测是非常困难的。数据挖掘是一种新兴的信息处理技术, 在信息的利用和提取中发挥着日益重要的作用。论文从数据挖掘的概念和本质出发, 阐述了数据挖掘的各种技术方法, 深入分析了数据挖掘技术在专利地图开发过程中的应用, 并对专利情报数据挖掘的方法和流程进行了探讨。

关键字: 专利分析; 专利地图; 数据挖掘

1 引言

专利作为技术信息最有效的载体, 囊括了全球 90% 以上的最新技术情况, 相比一般技术刊物所提供的信息早 5~6 年, 而且内容翔实准确^[1]。因此专利数据的分析早就被认为是一种评价技术变化中各个方面的方法, 大多数研究已经使用专利统计作为研究技术发展和经济增长之间关系的一种工具, 或者在国内和国际的条件下估计研究和革新的程序。专利分析已经成为一个用于分析公司政策研究、发展、估计竞争者的技术实力和缺陷, 以及开发市场的基础性研究工作。专利地图在专利信息利用中起到承上启下的重要作用, 承上是指将检索到的专利信息, 经过整理、加工、综合和归纳, 以数据的形式归入一张图表中, 可

供定量分析和定性分析之用; 启下是指通过对专利地图的对比、分析和研究, 可做出预测和判断, 从而得到可利用的技术水平、动态、发展趋势等情报, 为企业制定经营战略、专利战略、选定开发目标等服务。专利地图作为一种搜集、整理和利用专利信息的工具对企业有重要意义。

由于专利地图所涉及的信息量非常大, 整理起来异常烦琐。开发过程中必不可少的要用到数据挖掘技术。数据挖掘(Data Mining, DM)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中, 提取隐含在其中的、人们事先不知道、但又是潜在有用的信息和知识的过程。数据挖掘技术是 20 世纪末刚刚兴起的智能分析技术, 由于其所具有的广阔应用前

景而备受关注,作为数据库与数据仓库研究与应用中的一个新兴而富有前途的领域,可以从数据库中自动抽取或发现出有用的知识模式。运用数据挖掘技术对现有专利信息进行分析,以了解技术发展的生命周期或所处阶段、专利数据的地域分布与竞争对手的分布,可以减少重复研究和无效劳动,使投资行为更加合理主动。

2 数据挖掘技术

2.1 数据挖掘的定义

数据挖掘是一种利用各种分析工具建构数据分析模型,从而在大型的数据库(或数据仓库)中提取人们感兴趣的知识的过^[2]程。这些知识是隐含的、事先未知的、潜在有用的信息,提取的知识一般可以表达为概念、规则、规律、模式、约束、可视化等形式。这些表达形式蕴含了数据库中一组对象之间的特定关系,揭示出一些有用的信息,为科学研究、经营决策、市场策划、经济预测、工业控制等提供依据。

2.2 数据挖掘技术产生的背景

伴随数据库技术的广泛应用,企业管理中积累了大量的有用数据,包括市场、客户、供货商、竞争对手,以及未来趋势等重要信息,企业管理对数据处理技术要求不断提高。但是信息超载与非结构化,使得企业决策组织无法有效利用现存的信息,甚至使决策行为产生混乱与误用。基于总体假定进行推断和检验的传统的统计分析方法对这些海量、异构、分散数据的处理已显得无能为力,显露出很大的局限性。因此,如何从这些复杂的专利文献数据中以智能化的操作方式深入分析其中隐含的规律,如何发现、提取这些知识并加以利用就成了当务之急。

计算机技术的快速发展为专利情报分析提供了极大的便利条件,产生了数据挖掘等当今信息科学的前沿技术和管理方法,促使专利情报分析方法向自动化、智能化、可视化的方向发展。

2.3 专利情报数据挖掘的优势和发展现状

与传统的基于统计的专利情报分析法(如:查询、报表、联机应用分析等)相比,专利情报数据挖掘方法具有明显的优势。

1)方法优势。数据挖掘与传统的统计分析方法的本质区别在于统计是根据随机性的观测样本数据以及

问题的条件和假定,对未知事务做出以概率形式表述的推断;而数据挖掘则是在没有明确假设的前提下去挖掘信息、发现知识。与统计相比,数据挖掘工具处理大量的实际数据更有优势,无须专业的统计背景也能使用数据挖掘的工具,而且,数据挖掘从大型数据库提取所需数据,利用专属计算机软件进行分析,更能满足企业的需求。此外,从理论的角度来看,数据挖掘与统计也不同,其目的在于方便企业的末端使用者应用,而非为统计学家提供检验工具。

2)成果优势。数据挖掘技术同统计分析方法相比较,挖掘算法的研究成果无论从数量上还是实用性上都具有巨大优势。例如根据挖掘任务,数据挖掘算法包括分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等。近几年来,自然语言理解、语义关联分析、词频分布统计、语料学研究等可以用于进行情报分析的技术方法和工具,已经成为数据挖掘算法的重要研究方向。因此,无论从体系结构上,还是从具体方法上,数据挖掘算法都能够很好地融合到专利情报分析系统中去。

3)发展现状。目前,大多数国内学者对专利情报的分析仍停留在对数量特征的统计分析上,并且对于专利数据中的各种偏差性影响几乎都没有考虑,而国外学者则多利用数据挖掘的理论和数据可视化手段,正在研究专利信息的自动分析方法^[3]。总之,对于专利信息的利用,国内目前基本上仍处于定性分析和信息的统计计算管理水平上,尚不能从内容上自动挖掘专利数据中的知识规律。

3 专利情报数据挖掘的方法与流程

数据挖掘的产生并不是为了替代传统的统计分析技术,相反,对于统计分析方法,数据挖掘主要体现在利用统计和人工智能技术进行高级多元统计方法的应用上,是对这些方法的拓展和深化。因此,本文提出以数据挖掘为主、以数理统计为辅的两者结合的专利情报数据挖掘方法。专利情报数据挖掘的任务就是从专利数据中发现模式。数据挖掘的模式主要有关联规则、分类、聚类、序列模式等。与之对应的分析方法包括关联分析、分类分析、聚类分析、模式识别、可视化方法、人工神经网络等等。其中,关联分析是本文专利情报数据挖掘方法研究的重点。

在现有研究的基础上,根据实际需要,可将专利情报数据挖掘流程划分为以下四大模块:数据获取与数

据预处理、数据挖掘与监测、数据可视化以及情报分析与整合。其中,后三项统称为数据挖掘分析。

1)数据获取与数据预处理:根据专利情报分析工作的目的,确定解决问题的性质和数据挖掘的目标,明确科技监测的需求,初步选取监测对象。利用Spider技术,对下载服务器下达获取指令,服务器将相关数据进行获取,形成本地化数据库;对于获取的数据,采用数据预处理技术从目标数据集中剔除错误数据和冗余数据,去除噪声及无关数据,进行有效数据提取,经数据集成后,形成特定领域的情报监测数据集,为监测分析做准备。

2)数据挖掘与监测:结合监测需求,运用技术预测、自然语言处理、信息萃取、知识发现等工具,依据特定的数据挖掘算法(如关联分析、模糊聚类、技术组(群)自动识别、关键技术识别、自然语言处理等),在数据库中提取数据模式。

3)数据可视化:运用一定的方法将提取出的数据模式表达成某种易于理解的、直观的知识或模式(图、表等),对监测数据集和所拥有的本地化数据库进行初步分析,形成监测分析结果报告,提交给情报分析人员。

4)情报分析与整合:情报人员结合自己的专业知识,对监测分析结果进行系统、深层次地分析,在相关持续性投入技术、破坏性技术和新的研究方向等领域形成情报信息的监测分析报告,提交给该领域相关方面的专家。由专家凭借自己的知识和经验,对分析报告进行评估,再将评估意见反馈给情报人员,对监测报告进行修订,形成最终报告。由专家的专业知识来保证情报分析的有效性和可靠性。

4 相关的数据挖掘技术

4.1 时间序列分析

时间序列分析是数据挖掘的一个重要研究分支。很多情况下,所要分析的数据包含有一个时间维度,具有一个或多个依赖于时间属性的预测应用。传统的时间序列模型可分为自回归模型、滑动平均模型及自回归滑动平均混合模型三种。

4.2 神经网络技术

神经网络技术通过模仿大脑神经元的工作机制对系统历史、经验的数据进行学习,建立研究系统的等价模型^[4]。Kolmogorov连续性定理为神经网络奠定了坚实的理论基础;它证明了存在一个三层网络,其隐单元输出函数为非线性函数,输入及输出单元函数为线

性函数,此网络的总输入输出关系可以逼近任意一个非线性函数。因为任何一个时间序列都可以看成一个由非线性机制确定的输入输出系统,所以Kolmogorov定理从数学上保证了用神经网络对时间序列预测的可行性。

4.3 文本数据挖掘技术

文本是存储和交换信息最自然的方式,如何从纷繁芜杂的文本中发现有效的信息是信息处理的一大目标。传统的做法是对信息进行人工分类,并加以组织和整理,但是人工方法耗费大量的人力、物力和精力,而且分类结果一致性不高,因此文本挖掘具有非常重要的现实意义。文本挖掘是指将数据挖掘技术应用在大量的文本集合上,发现其中隐含知识的过程^[5];它是近几年来数据挖掘领域的一个新兴分支。文本挖掘是数据挖掘和信息检索两门学科的交叉,与传统数据挖掘技术的区别在于文本数据不同于传统数据,文本数据没有特定结构;因此,文本挖掘技术既采用了许多传统数据挖掘技术,又有自身的新特征。

利用自回归模型、灰色模型和广义神经网络模型,对专利基础数据进行处理,分析年、月专利数据波动周期,预测下一时间段的专利申请量,为专利预警机制打下基础。由于国际分类号复杂繁多,人工分类专利文档往往不准确,所以可以采用文本挖掘技术,利用KNN模型分类专利文档,依据分类模型的评估方法保证了分类的准确性和客观性,并在此过程中找出近似专利,供人工解读。

5 总结

专利地图分析软件对专利文献的分析可以由点及面的揭示技术信息^[6]。通过对单件专利的技术文献分析,可获得目前所遇问题的解决方法,获得技术发展新创意,回避现有专利陷阱。以时间为主轴,分析每年全球最新核准专利,可以快速反映新技术动向;以专利权人分析为主轴,可以了解该专利权人的技术与市场布局;以技术为主轴,时间、专利权人交叉分析,可以掌握技术趋势,对某一技术领域做出综合判断。通过专利地图分析,可以得到相关技术领域竞争者的如下信息:对手的研发活动年期分析、专利的年龄分析、专利的自我引证次数分析、被其他公司引证总引证次数分析和引证率分析、技术独立性分析、主要竞争者专利相互引证分析、重要公司的专利排行榜分析、重要公司历年专利件数分析。

专利地图一方面可以作为经营管理的重要信息,更重要的是还可以进一步归纳出每一个专利的技术及功效类别,通过分析这些图表隐含的与技术研发方向相关的信息,可以了解特定技术的动向,并进一步预测技术的未来走向。从以上角度来看,分析结果的正确性和可用性对于企业技术取向和产品的研发都有着至关重要的作用。数据挖掘就成为了目前最先进的专利情报分析方法。

References(参考文献)

- [1] LIU Ping, WU Xin-yin, QI Chang-wen. Patent Management Map Analysis on Technology of Laser Information Memory in Our Country. ELECTRONICS INTELLECTUAL PROPERTY, 2004 (3) : 28-32.
刘平, 吴新银, 戚昌文. 激光信息存储技术专利管理图分析[J]. 电子知识产权, 2004 (3) : 28-32.
- [2] HU Kan, XIA Shao-wei. Large Data Warehouse based Data Mining: a Survey. Journal Of Software, 1998,9(1).
胡侃,夏绍玮.基于大型数据仓库的数据挖掘:研究综述[J].软件学报, 1998, 9(1).
- [3] WU Xin-yin. Study on Application of Patent Map in Patent Strategy of Enterprise. Huzhong University of Science and Technology, 2004.
吴新银. 专利地图在企业专利战略中的应用[D]: [硕士学位论文]. 武汉: 华中科技大学, 2004.
- [4] ZHANG Zhi-qiang, ZHOU Li-zhu. Web Database Techniques: A Survey. Computer Science, 2001, 28(10):1-6.
张志强,周立柱,等.Web数据库技术简述[J]. 计算机科学, 2001, 28(10): 1-6
- [5] SHI Bing, ZHENG Yan-feng. Data Mining Technique of Information. Searching Journal of the China Society for Scientific and Technical Information, 199, 9, 18(s):103-106.
石冰, 郑燕峰.信息检索中的数据挖掘技术[J].情报学报, 199, 9, 18(s):103-106.
- [6] WU Xin-yin, LIU Ping. The Initial Probing into Patent Map. R&d Management, 2003 (10): 88-92.
吴新银, 刘平. 专利地图研究初探[J]. 研究与发展管理, 2003 (10) : 88-92.