基于知识发现的新一代网络浏览器

摘要:互联网时代的信息爆炸已使人们很难在海量信息中甄选出有价值的信息。同时，将海量信息萃取成知识的知识发现技术应运而生并迅速发展。当我们将知识发现的相关理论运用在浏览器上时，浏览器便会从单线信息相互链接向多样信息链接多样信息转变，这时我们发现，知识发现技术能够使我们获得“额外”的信息——通过所有数据获得的信息大于部分相加的和。而这种新一代浏览器，必将引发一场网络探索方式的变革。

关键词：信息爆炸 知识发现 浏览器 变革

0.引言

因特网克服了传统的时间和空间障碍，使得信息的采集、传播的速度和规模达到空前的水平，实现了全球的信息共享与交互，将世界更进一步地联接为一体。但与之俱来的“副作用”是：汹涌而来的信息有时使人无所适从，从浩如烟海的信息海洋中迅速而准确地获取自己最需要的信息，变得非常困难。面对信息爆炸，信息产业界迫切需要将海量数据转换成有用的信息。而需求是发明之母，近几年以数据挖掘为核心的知识发现技术引起了信息产业界的极大关注。另外，网络浏览器变得更加快捷、易于操作、辅助功能多样之后却始终不能带给用户更多的收获。当网络浏览器在瓶颈期遇到了知识发现技术，一种全新的网络探索方式便产生了。

目前，世界上对数据挖掘技术的应用已经初具规模。数据挖掘软件一般可分为两类：企业型数据挖掘软件可以提供多种数据挖掘算法，并能解决多种应用问题，如 IBM 的 Intelligent Miner 和 SAS Enterprise Miner 等；小型数据挖掘工具则是针对低端、低消费的用户，或是为解决特定的额应用问题提供特定的解决方案，如 Oracle 公司的 Darwin，Insightful 公司的 Insightful Miner 等。而基于知识发现的全新网络探索方式的研究也已初具雏形，如 Microsoft Live Labs 的 Gary Flake 宣布了基于 SeaDragon 和 Deep Zoom 技术的 Pivot 应用。Pivot 的浏览器基于 Trident 引擎，除了基本的浏览功能外，Pivot 最主要的就是允许用户自由筛选大量的可视化信息。Pivot 支持可视化收藏夹中网站链接，用户可在左侧的筛选面板中选择自己需要的条件，右侧的结果界面即可流畅切换至新的筛选结果（基于 Deep Zoom 深度缩放技术），并可自由缩放任何对象。但就目前版本而言，该软件还并不建议用户使用，因为它占用 CPU 很高，非常容易造成机器假死，分类方面也需要完善。

1.正文

1.1 核心思路

1. 1. 1 知识发现

知识发现(Knowledge Discovery in Database, KDD)，是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。它将信息转化为知识。知识发现过程由以下三个阶段组成：数据准备，2 数据挖掘，3 结果表达和解释。

数据挖掘的任务有关联分析、聚类分析、分类分析、异常分析、特异群组分析和演变分析，等等。在此基础上，知识发现的基本任务有数据分类、数据聚类、衰退和预报、关联和相关性、顺序发现、描述和辨别、时间序列分析等。

典型的基于算法的知识发现技术包括：或然性和最大可能性估计的贝叶斯理论 、衰退分析、最近邻、决策树、K 一方法聚类、关联规则挖掘 、Web 和搜索引擎、数据仓库和联机分析处理(On—line Analytical Processing，OLAP) 、神经网络、遗传算法、模糊分类和聚类、粗糙分类和规则归纳等。

这里介绍一种基于可视化的方法。基于可视化方法是在图形学、科学可视化和信息可视化等领域发展起来的，包括：

①几何投射技术。是指通过使用基本的组成分析、因素分析、多维度缩放比例来发现多维数据集的有趣投影。

②基于图标技术。是指将每个多维数据项映射为图形、色彩或其他图标来改进对数据和模式的表达。

③面向像素的技术。其中每个属性只由一个有色像素表示，或者属性取值范围映射为一个固定的彩色图。

④ 层次技术。指细分多维空间，并用层次方式给出子空间。

⑤基于图表技术。是指通过使用查询语言和抽取技术以图表形式有效给出数据。

⑥ 混合技术。是指将上述两种或多种技术合并到一起的技术。[1]

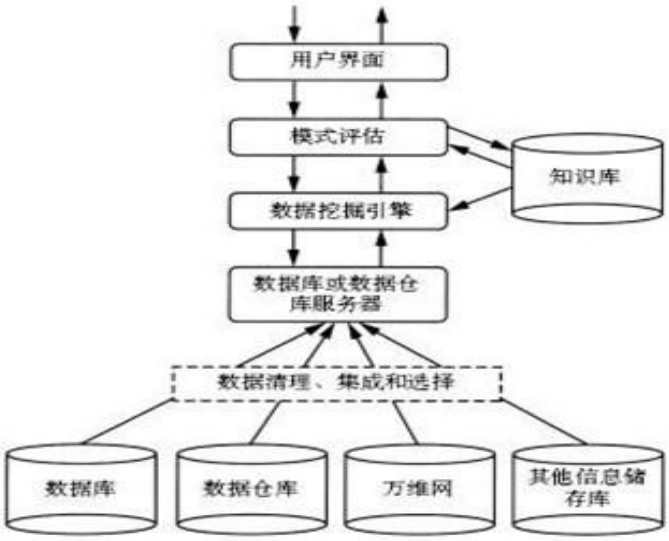

图 1

1.1.2 将知识发现技术应用于网络浏览器

网页浏览时，我们所面对的数据数量非常庞大，因此我们需要思考如何分析这些数据，揭开隐含于其中的规律和趋势。我们所熟悉的浏览器几乎都是线性的，但事实上网络并不是一页接着一页的线性旅程。应用了知识发现技术的浏览器，通过多维度的筛选能够呈现不同网页之间的联系。这样，我们通过所有数据得到的信息便大于部分相加的总和。

将知识发现技术应用于网络浏览器，应做到以下几点：

可以系统化的对用户积累的所有数据进行多样化分类，并建立庞大的数据库；

能够实现知识发现的基本任务，即数据分类、数据聚类、衰退和预报、关联和相关性、顺序发现、描述和辨别、时间序列分析等。通过这些任务，能够实现网页、表格、数据、摘要、图像等的多种分析，从而将数据信息转化为知识。这是新一代浏览器最核心最重要的功能。

能够将一切信息可视化、图像化，而不是简单的字符。并且将重新整合后的信息以最简洁的图像的形式展现在用户面前（基于 Deep Zoom 深度缩放技术）。

应用举例：Pivot

Gary Flake 在 TED2010 上演示了微软开发的 Pivot 软件。

Gary 向观众演示了用 Pivot 来分析一个典型的死亡率图表。一开始按照年龄分组分析死亡原因时，他得到一条曲线。曲线的中部表明，人们最活跃的年龄也是死亡率最高的时候。如果根据病痛死亡原因重组数据，可以看到循环系统疾病和癌症是致命的主要病症；如果按照年龄过滤（譬如 40 岁以下），可以看到意外事故是最致命的原因。如果进一步挖掘，还可以发现这一条规律尤其对男性适用。

Pivot 的作用在这一直观的演示中得以凸显。通过这种方式查看信息，数据像是在鲜活的信息资料图片中游走。不仅对原始数据可以这样做，对文字或图片等内容也可以。这是一种介于搜索和浏览之间的使用信息的方式。

之后，通过将页面压缩为一个小小的摘要（摘要包含了简介，且用一个图标来显示它来自的专业领域），Gary 还演示了用 Pivot 分析维基百科页面。他选取

了前 500 个最受欢迎的页面，但即使在这些有限的页面中，Pivot 依然可以做很多事情。例如我们可以得知维基百科上最流行的是什么。当你选择“政府”，你可以看到在维基百科类别中最常对应的是《时代》周刊年度风云人物。这一点很重要，因为这是一项不属于任何一个维基网页所载述的内容。只有退后几步俯瞰全局才有可能看得透彻。

图 2

1.2 产品可行性及市场前景分析

1.2.1 可行性

•技术方面 知识发现技术的发展已有一段时间，其中统计学、决策树理论、类神经网络、规则归纳法等都以比较成熟，目前已经可以支持大部分数据分析的需要。而将知识发现技术应用到浏览器上的其他技术也已相对成熟，如图形可视化技术、深度缩放技术等。因此技术方面，该类浏览器的生产是可行的。

• 硬件方面 由于该类浏览器为了使用的流畅性、统计分类的快速性、用户体验的舒适性等，必将采用多种非常占用 CPU 资源的技术，就目前的市场来看，此类软件还不能快速在市场上普及。但随着时间的推移，以计算机行业的“摩尔定律”来看，此类软件在未来几年至十几年的时间内，计算机硬件一定会比今日有很大的飞跃，到那时，新一代浏览器的普及还是相当可行的。

总之，无论在技术方面还是硬件方面，基于知识发现的新一代网络浏览器的生产都是相当可行的。

1.2.2 市场前景

传统浏览器在浏览模式不改变的情况下，很难有其他质的飞跃，而走向新一代网络浏览方式则是大势所趋。作为网上冲浪的必经之路——浏览器的使用对象不是某一群体，而是所有上网的用户，这是一个相当巨大的群体。而在二十一世纪这个全球信息化的时代，使用网络的用户一定会不断快速增长。目前，微软还在研发试用阶段的 Pivot 软件就是此类新一代浏览器拥有美好市场前景的有力例证。

总之，基于知识发现的新一代网络浏览器绝对拥有广阔而光明的市场前景。

参考文献

[1] 李雄飞,李军.《数据挖掘与知识发现》.高等教育出版社.