美国专利引证可视化系统的设计与实现

刘玉琴[1],彭茂祥[2]
LIU Yuqin[1], PENG Maoxiang[2]
1.中国科学技术信息研究所,北京 100038
2.北京理工大学 管理与经济学院,北京 100081
1.Institute of Scientific and Technical Information of China, Beijing 100038, China
2.School of Management and Economics, Beijing University of Technology, Beijing 100081, China

LIU Yuqin, PENG Maoxiang. Design and implementation of visualization system of US patent reference. Computer Engineering and Applications, 2012, 48(22):34-40.

Abstract: By analyzing the disadvantages of the current patent reference analysis tools, the paper introduces the basic framework, key technologies, algorithms and realization of patent reference search and visualization system. The difference of visual analysis method between this paper and the previous is proposed.
Key words: patent reference; patent analysis; visualization

摘 要:分析了目前专利引证分析工具现状与不足,介绍了美国专利引证可视化系统的总体结构、关键技术、主要算法以及可视化实现,指出了系统构建的可视化分析方法与以往分析方法的不同。
关键词:专利引证;专利分析;可视化
文章编号:1002-8331(2012)22-0034-07 **文献标识码:**A **中图分类号:**TP391

现代社会进入了信息社会,信息资源已成为现代社会中最重要的战略资源之一。专利作为一种特殊的信息和战略资源,在国家信息资源建设开发与利用中有着特殊的地位和作用。其中专利引证信息作为专利文献的重要组成部分,更是发挥着举足轻重的作用。通过专利的被引证数量统计可以发现特定技术领域中重要核心专利的分布,进而开展科技成果评价;通过跟踪专利的被引证信息可以发现竞争对手和潜在竞争对手及其技术研发策略;通过回溯专利的引证信息可以揭示专利的初始技术来源;通过专利间引证与被引证信息的综合可以揭示专利间相互联系、相互影响与相互促进的关系;通过分析引证专利和被引证专利所解决的技术问题与技术创新性可以揭示特定技术领域内的知识流动、技术扩散、技术融合的路径。专利引证信息的分析在所有的专利信息分析中占有非常重要的地位。

然而,在面对海量的专利信息资源时,如何高效、准确地开展专利分析工作,向专利分析人员提出了新的挑战。可以说,专利信息分析挖掘方法、分析挖掘工具的合理使用决定了专利信息获取的准确性和有效性,以至于决定了专利分析人员的最终的信息分析水平、效率以及信息分析质量和效益。本文在分析现有专利引证分析工具及其不足的基础上,主要讨论使用信息可视化方法进行美国专利的引证分析。

1 专利引证分析工具现状与不足

国外对于引证分析的研究由来已久,用于专利引证分析的工具也很多。概括起来这些工具可分为两大类:具有引证分析功能的专利信息服务平台和基于引证分析方法的软件工具。

在专利信息服务平台方面,以美国汤姆森科技的 Innovation、Aureka 专利创新平台,美国 Dialog 公

基金项目:国家科技重大专项(No.2009ZX01030-003-003-5)。
作者简介:刘玉琴(1979—),男,高级工程师,在站博士后,主要研究方向:专利分析、信息可视化;彭茂祥(1970—),男,博士生,主要研究方向:技术预测、专利分析。E-mail:liuyuqin2004@126.com
收稿日期:2011-12-14 **修回日期:**2012-03-26
DOI:10.3778/j.issn.1002-8331.2012.22.007

司的Innography专利检索分析平台，LexisNexis公司的TotalPatent，法国Orbit公司的Questel专利检索分析平台，韩国世界知识产权检索株氏会社WIPS专利检索分析平台为典型[1]，这些服务平台拥有丰富的数据资源、先进的数据挖掘技术、可视化技术以及领先的行业经验，将“数据、检索、分析、服务”进行捆绑，在为国内相关企业提供信息服务时，往往提出较高的价格，限制了中小企业用户和个人用户的使用。

在引证分析的软件工具方面，也以国外的工具和产品居多。这些工具最初设计是进行论文引证分析的，配合其他辅助工具也可以进行专利的引证分析，其中典型的有美国汤姆森科技HistCite[2]，美国Drexel大学陈超美的CiteSpace[3]，这些软件工具的特点是功能丰富，价格低，有些甚至免费，只是在进行专利分析时往往需要配合其他数据预处理工具进行数据的转换，有时甚至要编写代码进行数据的预处理，因此，在专利引证分析中的实际应用并不多。

无论是具有引证分析功能的专利信息服务平台还是基于引证分析方法的软件工具，在专利引证分析实现上，采用了数据统计列表、直方图、一般层次树结构、双曲树结构、社会网络进行分析结果的呈现，除了双曲树和社会网络分析的算法相对复杂以外，其他的分析呈现手段相对简单，只要有完备的引证数据即可实现，而这恰恰是国内专利分析工具所欠缺的。

同时，这些分析方法对于多级专利引证信息的揭示能力有限，以Innovation和Aureka的专利引证双曲树为例，通过其进行的专利引证分析较统计列表、直方图和一般层次树结构更直观，所表现的信息更加丰富，但被分析的引证级别一般只能到2级，其原因主要是专利引证数据随着引证级别的加深，数据成倍数上升，因此，利用它来进行技术跟踪还存在一定的局限。再以CiteSpace社会网络分析为例，其底层算法采用的是开源软件，对于算法本身没有改进，在引证专利数量较多时，网络节点位置计算时间成本上升，可视化结果的可读性降低。

国内对专利引证分析的研究一方面侧重于引证分析理论、概念方法或国外分析工具的介绍[2,4-5]，另一方面侧重于应用引证分析方法进行特定领域的实证研究[6-9]。少数研究者从可视化角度进行专利引证信息的分析与系统研究，如张兆锋等基于可视化技术设计的专利引证分析工具[10]。该工具偏重于微观层面的专利与专利间的引证可视化，可视化的结果以一般的层次树结构进行显示，可视化结果相对单一，而且主要针对固定的引证数据进行分析，实时性差，在分析海量专利引证关系时，可视化结果不易阅读。

现实应用中，由于美国的专利引证信息在各国引证数据中最为完备，且免费开放，在专利分析领域最为常见。为此，针对目前国外专利引证分析工具与专利数据进行绑定、服务价格高、多级专利引证信息揭示能力有限，国内专利引证分析工具缺乏，而且分析结果不具实时性、分析方法单一以及在海量专利引证分析上的不足等问题，本文提出了一种针对美国专利引证信息的可视化系统设计与实现，可提供实时的专利引证信息，并将引证分析结果以多种可视化图形方式展示。

2　美国专利引证信息可视化系统

2.1　系统结构

本文构建的美国专利引证可视化系统主要由专利引证信息的采集、清洗转换和可视化分析几个功能组成。专利引证信息主要来源于美国专利商标局的官方检索平台。系统结构如图1所示。

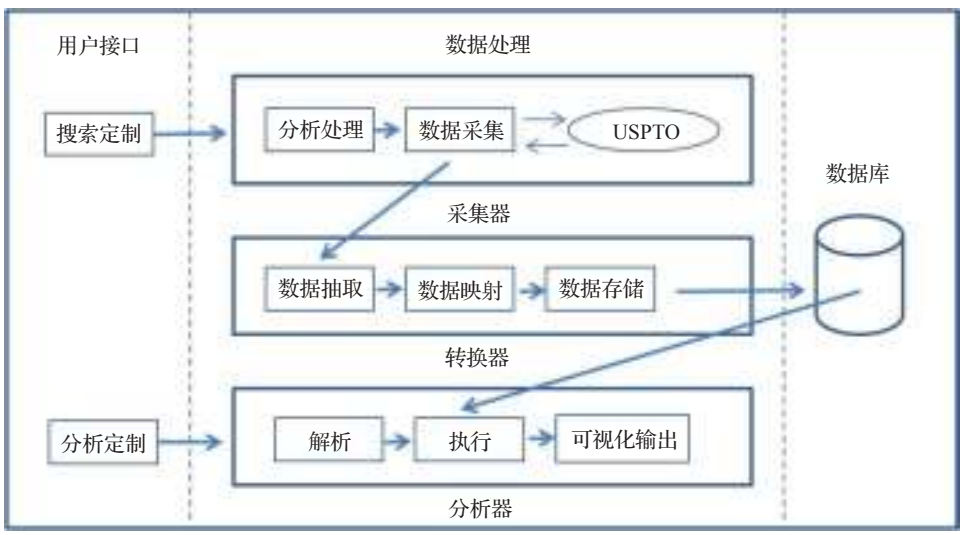

图1　美国专利引证可视化系统结构

在专利引证采集过程中,用户输入单件或多件专利的专利号后,系统采集模块启动两个主要的采集线程,一个线程以深度优先算法对输入专利的引证数据进行采集,另一个线程对线程一采集到的引证专利的申请人、发明人、申请时间、摘要、权利要求等著录项目进行补充采集,著录项目一方面用来帮助使用者进行引证专利内容的理解,一方面用来作为更深级别引证信息采集的定向依据。采集模块采集的信息均为HTML,要由清晰转换模块进行内容抽取,以获取规范的著录信息,清洗模块调用存在数据库中的正则表达式集合进行对应内容的抽取,把抽取出的规范数据存放到数据库,并且把信息组合成专利节点以便在可视化模块中使用。采集模块不涉及算法,但需要对美国专利网站的页面结构进行详细的解读,编写出正确的正则表达式。可视化模块对清洗后的数据进行可视化呈现,首先构建专利节点或专利发明主体节点集合,将节点作为双曲树的树节点或者网络图的图节点,根据用户的分析需求,调用双曲树算法或Fruchterman-Reingold布局算法进行节点的空间位置计算,最后由可视化渲染器进行节点的外观显示。

2.2 关键技术

(1)引证信息的实时定向搜索

针对多级专利的引证级别多、引证数量大,一次性搜索、显示的时间成本和空间成本大等问题,系统在进行单件专利的多级引证分析时采用实时定向搜索的策略,实现的流程如图2所示。

首先,在美国专利引证数据库中进行待分析专利的引证信息搜索,返回一级引证专利号集合。然后,以这些专利号为参数向美国专利基础数据库发送搜索请求,搜索这些专利号对应专利的著录项目信息。在进行二级搜索时,用户根据一级引证专利的著录项目,如技术类别、申请人、申请日等选择进行二级引证搜索的专利号作为定向依据,发送二级引证信息的搜索指令。依次类推,重复第N级专利引证信息的定向与搜索。通过这种方式,使得用户在进行引证信息分析时,把分析重点放在其关注的技术领域和专利所有者上,从而节省搜索和分析的时间。

(2)引证信息的可视化表示

在针对被分析专利为单件专利的情况,本文采用树形结构和网络结构进行可视化显示,将专利作为树形结构图和网络结构图中的节点,用节点间带有箭头的连线指示专利间的引证关系。单件专利的引证关系可视化显示与实时定向搜索紧密结合,使用户将重点放在其关注的专利引证信息上。

在针对被分析专利为多件专利的情况(比如,同

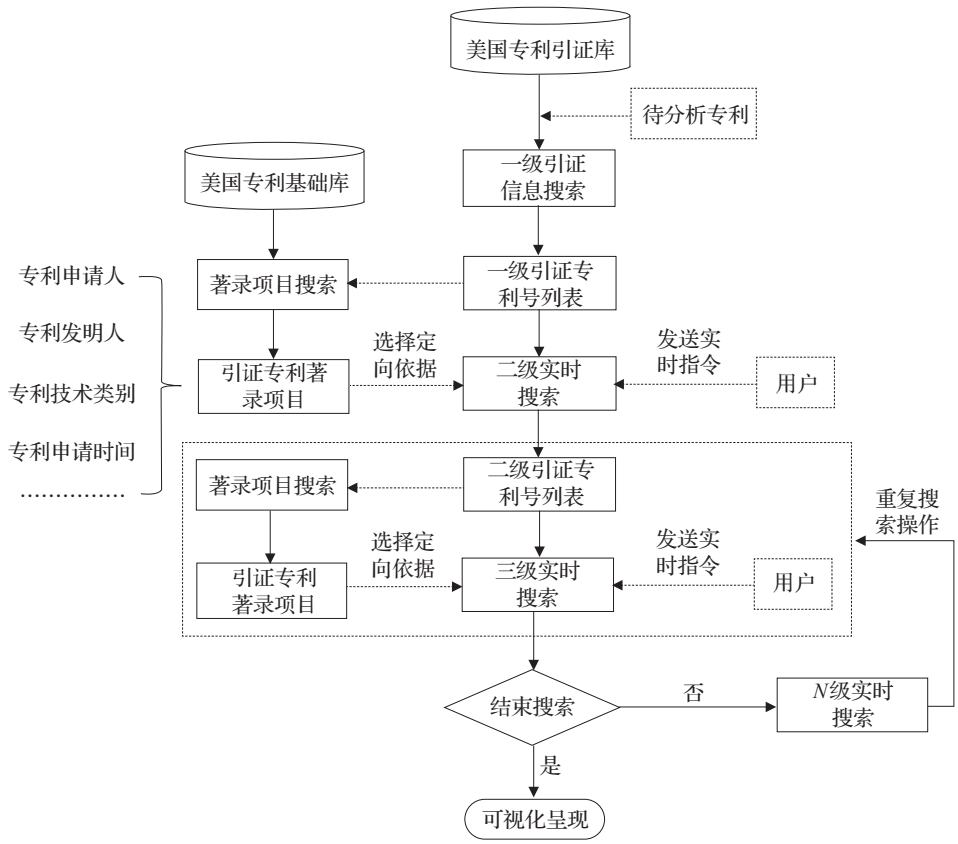

图2 单件专利多级引证实时定向搜索流程图

一申请人的专利),本文采用网络结构进行可视化显示。如果被分析的专利与搜索的引证专利数量不大,则将专利作为网络结构图中的节点,用节点间带有箭头的连线指示专利间的引证关系;如果数量较大,则进行显示内容的转换,进行专利节点的归类合并,具体过程如图3所示。

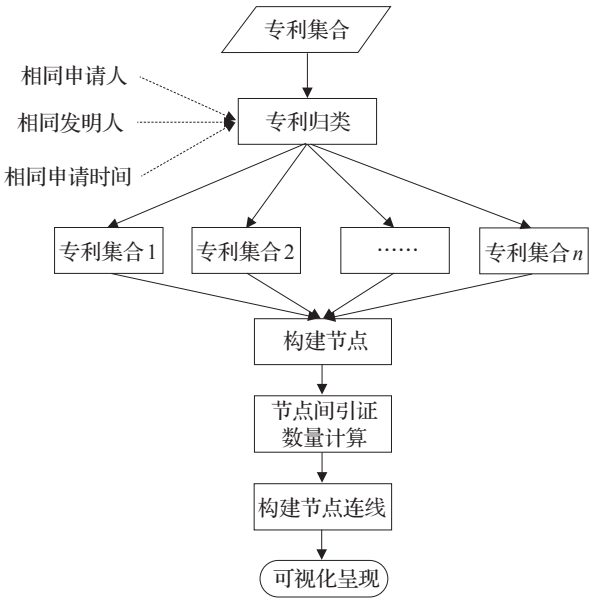

图3 发明主体年代引证关系构建示意图

首先,把具有相同的发明主体(申请人或发明人)和申请时间的专利归为一类,以每一类别(即发明主体×年代)作为网络节点;然后,计算每一个类别下专利被其他类别下专利引用的数量和作为这一类的被引用数量,调整节点大小与该类别被引用数量成正比。最后,用节点间带有箭头的连线指示类别间的引证关系,连线的粗细与引用数量成正比。通过这种变化,可以缩小引证数据的规模,同时,也可以揭示专利技术在不同年代、不同发明主体之间的演化关系。

2.3 主要算法

本文在专利引证可视化上采用了Hyperbolic Tree双曲树算法进行专利引证信息的树形结构呈现,采用Fruchterman-Reingold复杂网络算法进行专利引证信息的网络结构呈现。

2.3.1 Hyperbolic Tree

(1)算法原理

双曲树算法由美国Xerox Palo Alto研究中心提出[11],其主要原理是将树结构在双曲空间进行布局,然后映射到欧式空间的庞莱卡圆盘进行显示,映射的示意图如图4所示。欧式空间中两个相同大小的区域离庞莱卡圆盘中心越近,在双曲空间中所占用的空间越小;反之,双曲空间中两个大小相同的区域离原点越近在庞莱卡圆盘中所占用的空间越大。所以,当把使用者关注的树节点放到双曲空间的原点后,在欧式空间该节点显示在圆盘中心(电脑屏幕中心),而且占用的空间最大。

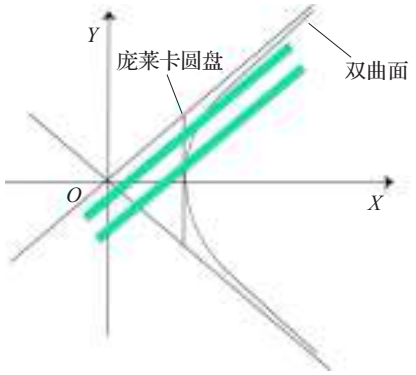

图4 双曲空间映射示意图

(2)技术实现

算法包括双曲空间树节点布局,双曲空间向欧式空间映射两个主要步骤。首先将树的根节点坐标设置为双曲平面的原点(0,0),然后把根节点的扇形区域平分给根节点的子节点,这样每个第二级节点都有自己的扇形区域,再把第二级子节点的每个扇形区域平分给其所拥有的第三级子节点,依次递归进行节点的分布;接下来采用庞莱卡投影把双曲空间中的点和线映射到欧式空间的庞莱卡圆盘上。具体技术实现步骤如下:

步骤1 定义复数类HTCoordinate(double x, double y)用以表示双曲空间中点的位置,复数的实部、虚部分别与双曲空间中点的横纵坐标对应,扇形类HTSector(HTCoordinate p1,HTCoordinate p2)用于表示双曲空间中的扇形区域。

步骤2 在双曲空间中对树结构中的每个节点进行布局,除了根节点布局在原点外,其他节点调用如下布局过程进行坐标的递归设置。

```
HTCoordinate w=parent.Coordinates(); //获取当前父节点坐标
double angle=sector.Angle(); //获取当前节点父节点所拥有的扇面
z.x=length*Math.Cos(angle); //获取当前节点相对于父节点的位置坐标
z.y=length*Math.Sin(angle);
z.Translate(w); //经过变换设置当前节点在双曲空间中的坐标
```

步骤3 将双曲空间点Z的坐标映射到欧式空间,映射规则如下:

```
x=Math.Round(z.x*(double)max.x)+org.x;; //max为庞卡莱圆盘的大小
```

y=-Math.Round(z.y*(double)max.y)+org.y; //org为庞卡莱圆盘中心的欧式空间坐标

2.3.2 Fruchterman-Reingold layout

(1)算法原理

Fruchterman-Reingold复杂网络布局算法由Fruchterman T M J和Reingold E M[12]提出,简称FR算法。该算法建立在粒子物理理论的基础上,将无向图中的节点模拟成原子,通过模拟原子间的力场来计算节点间的位置关系。算法通过考虑原子间引力和斥力的互相作用,计算节点的速度和加速度,节点的运动规律类似原子或者行星间的运动,系统最终进入一种动态平衡状态。

(2)算法改进

在FR算法的实现上,本文做了一些改进的措施:在计算引力时,预先设定一个引证数量阀值,只有那些连接线所代表的引证数量超过该阀值时,才计算连接线两端节点的引力,低于这个阀值的连接线不显示,也不计算其两端节点的引力。这样改进的益处在于:用户随时调节阀值,把那些明显的引证关系突显出来,同时加速算法本身的计算速度。

(3)技术实现

步骤1 首先对网络图中的每个节点坐标进行随机初始化。

步骤2 计算任意两点间$p1$、$p2$的斥力,并根据斥力大小设置平移,代码如下:

```
double dx=p1.X-p2.X;
double dy=p1.Y-p2.Y;
double dLength=Math.Sqrt((dx*dx)+(dy*dy)); //两点间距离
double force=(repulsion_constant*repulsion_constant)/dLength; //FR斥力公式 f_r=-k^2/r
double dx'=(dx/dLength)*force;double dy'=(dy/dLength)*force; //点的位置偏移量
```

步骤3 计算边e两端节点的引力,并根据引力大小设置平移,代码如下:

```
p1=e.start;p2=e.end;
double dx=p1.X-p2.X;
double dy=p1.Y-p2.Y;
double dLength=Math.Sqrt((dx*dx)+(dy*dy)); //两点间距离
double force=(dLength*dLength)/attraction_constant; //FR引力公式 f_a=r^2/k
double dx'=(dx/dLength)*force;double dy'=(dy/dLength)*force; //点的位置偏移量
```

步骤4 重复步骤2和步骤3,直至网络图的结构便于用户理解。

3 系统实现

3.1 功能简介

系统采用微软C#+WPF作为客户端开发工具,数据库采用SQL Server。功能模块主要分为引证信息实时搜索与清洗模块,引证信息可视化分析模块,PowerPoint演示文稿输出模块。图5、图6分别为系统搜索和可视化显示内容切换页面。

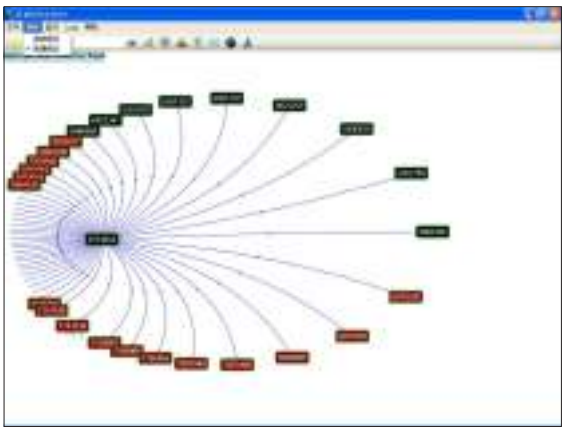

图5 系统搜索页面

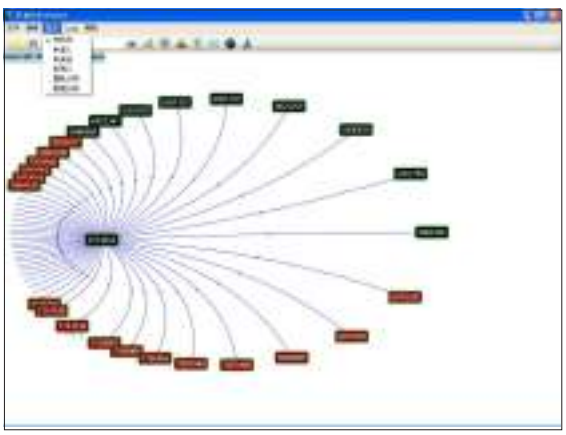

图6 系统显示内容切换页面

引证信息实时搜索与清洗模块主要实现专利引证信息的采集与数据清洗,可按照单件专利、多件专利分别设定搜索入口,同时,指定搜索专利的技术类别、申请人、发明人等,从而为分析模块提供数据基础;引证信息可视化分析模块则采用双曲树和社会网络分析分别实现单件专利多级引证树,单件专利引证网,多件专利引证网,专利发明主体年代引证网;Microsoft PowerPoint演示文稿输出模块将可视化分析结果直接以演示文稿的形式输出,提供演示接口。

3.2 可视化输出

(1)单件专利的引证可视化

图7、图8为系统分析单件专利引证的可视化输出,分别为单件专利多级引证树和单件专利引证网。

在单件专利多级引证树的可视化中,用节点与节点的文字标注专利的授权号、发明人、申请人、授权时间、国家分类号、美国分类号等信息;用箭头表示引用与被引用的关系;节点颜色用于区别引用与被引用关系;鼠标的拖动来更改关注专利的焦点。使用时,首先输入被分析专利,搜索该专利的一级引证和被引证专利,将其全部显示出来,示例图7中为6753654;在进行二级引证搜索时,按照用户指定的技术类别、申请人、发明人有针对性地把部分引证专利和被引证专利进行实时定向搜索后再采用双曲树算法进行可视化表示,使用户集中精力关注其感兴趣的信息,示例图7中选择了与被分析专利具有相同申请人且技术类别相同的专利7141934进行二级引证的搜索和可视化,按照相同的操作选择定向条件进行更深层级的引证信息搜索和可视化表示;最后,用户可就其感兴趣的信息进行节点显示内容的切换。

在单件专利引证网络可视化中,与单件专利多级引证树的可视化类似,同样用节点来表示专利的授权号、发明人、申请人、授权时间、国家分类号、美国分类号等信息,箭头表示引用与被引用的关系,不同之处在于其关注的内容强调被分析专利在整个引证网络中的地位以及其他与之有关联的专利之间的引证关系。在单件专利引证网络中,随着引证层级的增加,专利引证数量增加,可视化应用的FR算法计算时间成本增加,可视化结果的可读性降低。因此,只在引证数量较少的情况下应用。

(2)多件专利的可视化

图9、图10为系统分析多件专利引证的可视化输出,分别为多件专利引证网和发明主体年代引证网。与单件专利引证网相似,多件专利引证网只在引证数量较少的情况下应用,当引证数量较大时则采用发明主体年代引证网。

发明主体年代引证网是为了应对海量专利的引证关系设计的可视化方法,如本文2.2的关键技术分析所述,当专利数量较多时,很难再从专利间的引证关系探寻整体的技术演化,从发明主体年代间的引证关系上进行可视化,不但能够发现专利技术之间的演化,也大大缩小了数据处理的规模。在发明主体年代引证网络中,用节点大小表示发明主体当年专利被引用的总数量,用箭头表示哪些主体在什么年代引用了,箭头仍然表示引用与被引用关系,用节点文字标注发明主体名称及其在当年最多的专利技术类别。通过这种可视化表示可以从宏观上对发明主体专利的引证情况进行了解,定位主要的发明主

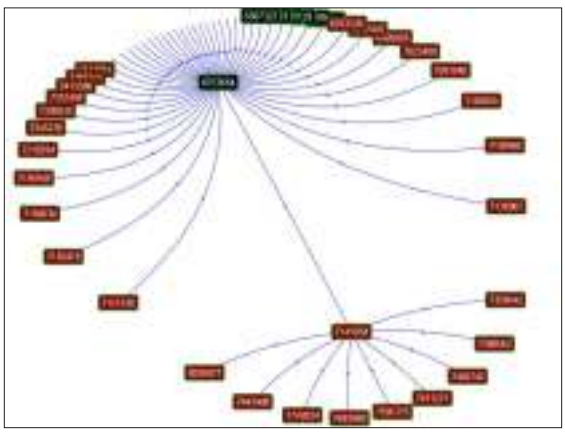

图7 单件专利多级引证树

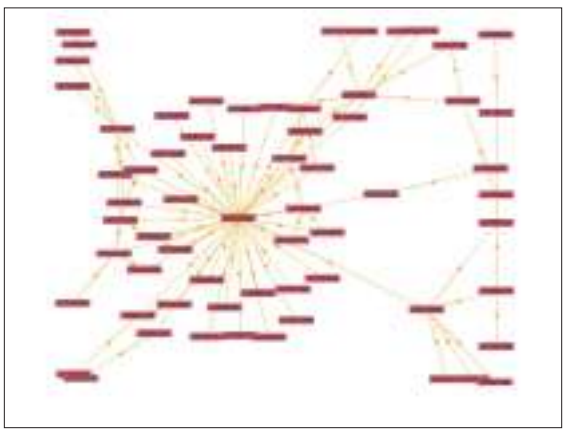

图8 单件专利引证网

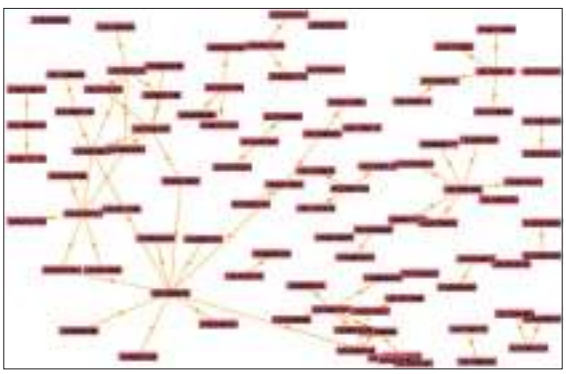

图9 多件专利引证网

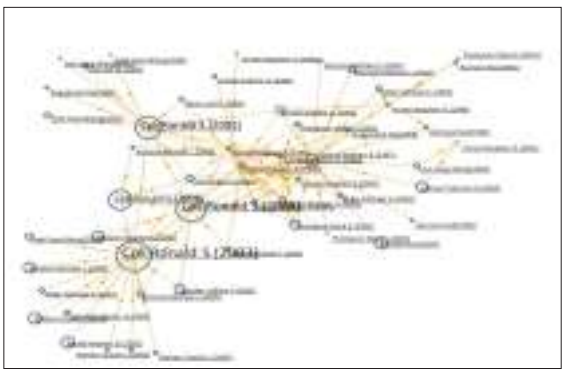

图10 专利发明人年代引证网

体和发明时间及其技术演化关系。

以苹果公司的美国授权专利为例,截止到2011年10月,该公司共有专利4 108件,图11是对这四千多件专利进行的网络可视化结果。在可视化图形计算过程中消耗大量的时间,可视化结果因为节点和连线过多而变得难以理解。图12是从申请机构×年代的角度构建引证关系,不但降低了网络图的规模,便于理解可视化结果,而且,可使用户快速地获知各个年代之间专利技术的引证关系,从而达到对苹果专利技术演化路径跟踪的目的。

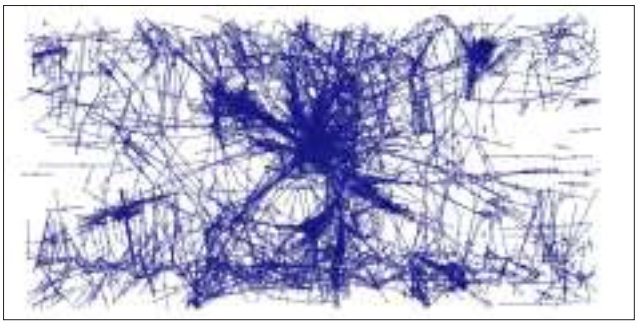

图11 苹果公司专利引证网

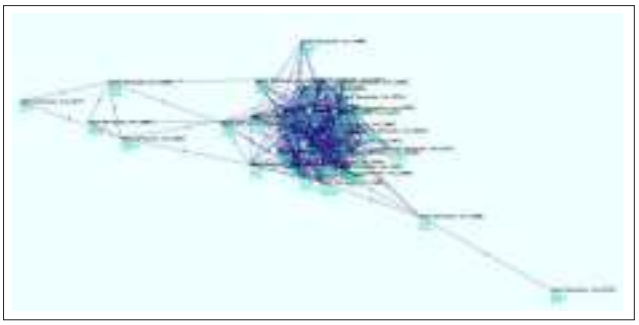

图12 苹果公司专利申请人年代引证网

4 结论

本文设计并实现了一个美国专利引证可视化系统,应用信息可视化技术对专利引证情况进行分析,分别实现了专利的引证树、引证网。与以往引证可视化分析工具不同,本文构建的引证树,可按照用户指定的条件进行定向搜索和有针对性的可视化分析,节省了分析的时间成本。同时,本文首次从专利发明主体的角度建立发明主体年代之间的引证网络,有利于发现海量专利间的技术演化路径。在将来的系统扩展中,还会考虑将欧洲、日本等专利局网站作为信息源,扩大专利引证信息分析范围。

参考文献:

[1] 30种常见的专利分析产品[EB/OL].[2011-11-08].http://blog.sciencenet.cn/home.php?mod=space&uid=394038&do=blog&id=425873.

[2] 李运景,侯汉清,裴新涌,等.引文编年可视化软件HistCite介绍与评价[J].图书情报工作,2006,50(12):135-138.

[3] 陈玉光,丁堃,刘盛博.基于CiteSpace II的专利知识可视化的实现机制及其应用[J].情报学报,2010,29(8):663-670.

[4] 孙巍,张学福.基于引文的信息检索可视化相关系统比较分析[J].情报理论与实践,2008,31(4):598-601.

[5] 苑彬成,方曙,刘清,等.国内外引文分析研究进展综述[J].情报科学,2010,28(1):147-153.

[6] 康宇航,苏敬勤.基于专利引文的技术跟踪系统——理论模型与工具开发[J].科学学与科学技术管理,2008(4):24-27.

[7] 李昌新,唐惠燕,陆芹英.引文计量与学术影响——用两种中文引文库评价农业科教系统学术地位的统计分析[J].农业图书情报学刊,2000(5):55-57.

[8] 贾玉英.《系统工程理论与实践》作者及引文的统计分析[J].农业图书情报学刊,2006,18(10):155-157.

[9] 王俊文,崔蒙,赵英凯.35中医药领域系统评价文献的引文分析[J].中医研究,2011,24(3):67-70.

[10] 张兆锋,桂婕,乔晓东,等.专利引证分析工具的设计与实现[J].数字图书馆论坛,2010(9):20-25.

[11] 窦长威,刘晨.层次信息可视化技术的一种实现方法[J].工程地质计算机应用,2007(2):11-15.

[12] Fruchterman T M J,Reingold E M.Graph drawing by force directed placement[J].Software Practice and Experience,1991,21(11):1129-1164.

(上接9页)

[80] Fischer S.Visuelle navigation mit parameter-modellen[D].Faculty of Technology,Bielefeld University,2006.

[81] Vardy A,Möller R.Biologically plausible visual homing methods based on optical flow techniques[J].Connection Science,Special Issue:Navigation,2005,17(1/2):47-89.

[82] Möller R,Lambrinos D.Insect strategies of visual homing in mobile robots[M]//Biorobotics-Methods and Applications.[S.l.]:AAAI Press/MIT Press,2001:37-66.

[83] Nister D,Stewenius H.Scalable recognition with a vocabulary tree[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2006:2161-2168.

[84] Shahbazi H,Zhang Hong.Application of locality sensitive hashing to real-time loop closure detection[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems,2011:1228-1233.

[85] Hornung A,Bennewitz M,Strasdat H.Efficient vision-based navigation learning about the influence of motion blur[J].Autonomous Robots,2010,29(2):137-149.

[86] López-Nicolás G,Sagüés C.Vision-based Exponential Stabilization of Mobile Robots[J].Autonomous Robots,2011,30(3):293-306.