# 专利引证分析工具的设计与实现*

□ 张兆锋 桂婕 乔晓东 朱礼军 李鹏／中国科学技术信息研究所 北京100038

摘要：专利引证分析在专利分析中的作用越来越大。文章介绍了一个专利引证分析工具的设计与实现，包括数据的获取、清洗、转换和生成图形，并进行了简单的应用分析。同时，介绍了一个开源的可视化工具包Prefuse，希望对想通过用可视化手段进行文献分析的研究人员提供有益的启示。

关键词：专利引证，信息可视化，Prefuse，数据挖掘
DOI：10.3772/j.issn.1673—2286.2010.09.005

## 1 引言

随着科学技术的迅速发展和企业间的竞争日趋激烈，竞争情报在企业战略、产品战略制定中的作用越来越明显。分析、研究和利用专利信息已成为获取竞争情报的一种重要手段。国内对专利分析的重视也逐步提高，但更多的专利分析及报告只是基于专利的外部特征所作的简单统计、排序等处理，而没有把专利资源的价值进行充分的挖掘利用。

引文分析一直是文献计量学中的重要内容，人们通过文献之间相互联系、相互影响与相互促进的关系，并对文献所承载的科学研究成果进行评价。借鉴引文分析的思想，专利引文分析已成为引文分析中的重要分支，也是充分挖掘专利信息的一种方式。专利引文分析即是利用各种数学与统计学的方法及比较、归纳、抽象、概括等逻辑方法，对专利文献引用参考文献的现象进行分析研究，从而揭示其数量特征和内在规律，并据此进行技术发展趋势的评价及预测[1]。

专利之间的相互引用与文献之间相互引用非常相似，专利引用分析方法是以前后专利之间的引用与被引用关系为基础，结合适当的方法，对专利间的相互引用现象和规律进行分析，以挖掘出某一企业或某一行业潜在的趋势和规律的一种专利定量分析方法[2]。利用专利引证分析，可以方便地了解某行业核心技术或某项技术的缺陷，可以认识潜在的竞争对手及其核心技术、专利保护策略和研究方向等，还可以评估自己公司的专利策略是否完善，遇到技术问题时，也可以寻求不同的解决方案。

国内对专利引文分析介绍较多，但大多数仅限于理论概念阶段，或是对国外软件的应用介绍。本文尝试介绍一个用于可视化开发的开源工具包Prefuse，及基于它的一个专利引证分析工具的设计与实现过程，为需要进行专利引证分析或进行专利深度挖掘工具开发的研究人员提供一种思路。

## 2 设计原理

### 2.1 信息可视化

信息可视化就是利用计算机支撑的、交互的、对抽象数据的可视表示，来增强人们对这些抽象信息的认识[3]。

信息可视化的内涵是将数据通过图形化、地理化形象真实地表现出来并且找到数据背后蕴含的信息。信息可视化相关技术能够实现对信息数据的分析和提取，然后以图形、图像、虚拟现实等易为人们所认识的方式展现原始数据间的复杂关系、潜在信息以及发展趋势，以使我们能够更好地利用所掌握的

* 本文得到国家科技部“十一五”科技支撑计划（项目编号：2006BAH03B03）、中国科学技术信息研究所重点工作项目（项目编号：2009KP01—7—1）、中国科学技术信息研究所2009年度预研基金项目（项目编号：YY—200906）等项目的资助。

信息资源。

本引证分析工具即采用信息可视化的原理，基于开源的Prefuse工具包把专利之间的引证关系以引证树的方式形象化地展示出来，可以让用户更方便地查看专利之间的前后引证关系及技术发展的脉络或某公司的专利布局情况。

### 2.2 Prefuse开源软件包

Prefuse是一种基于JAVA的开源的软件框架，可以用来创建具有交互性的信息可视化应用程序。Prefuse既可以创建独立的应用，也可以作为可视化组件嵌入在型应用程序中，或者制作成Applet小程序在网页上展示。它有如下的特色[4]：

(1) 它可以方便地处理和展示数据，如可以通过改变结点显示的位置、大小、形状、颜色等。

(2) 支持表结构、图结构、树结构数据的展示及基于这些数据的查询结果的展示。

(3) 可以对图的大小进行任意的缩放。

(4) 支持动态查询，交互性地过滤数据。

(5) 整合了有效的文本检索引擎。

(6) 支持多种视图样式，可进行预览视图和详细视图的切换。

(7) 支持类似SQL的查询语句，支持对Prefuse数据结构的查询及展示。

(8) 有友好的API函数帮助支持开发者创建自己想要的交互式展示功能。

Prefuse工具包是基于信息可视化参考模型进行设计的，这个模型把可视化的过程分成一系列独立的步骤。如图1：首先从原始数据（Source Data）通过数据转换符合展示建模要求的结构化数据（Data Tables），然后通过可视化映射把相应数据进行可视化的抽象建模（Visual Abstraction），最后再把抽象出来的模型通过视图转换展示出来。

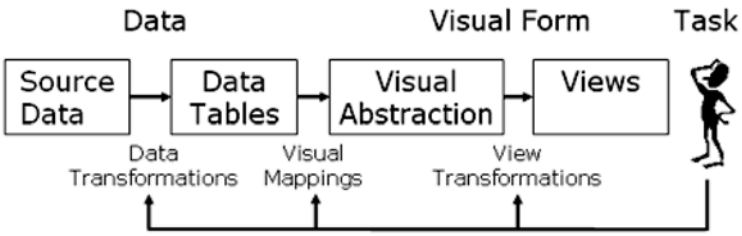


图1 信息可视化参考模型[5]

## 3 系统构建

### 3.1 数据流程

本文所采用的数据处理流程如图2：下载原始网页、抽取特征项、对数据进行清洗转换，最后形成符合图形展示条件后才可以生成的可视化图形。

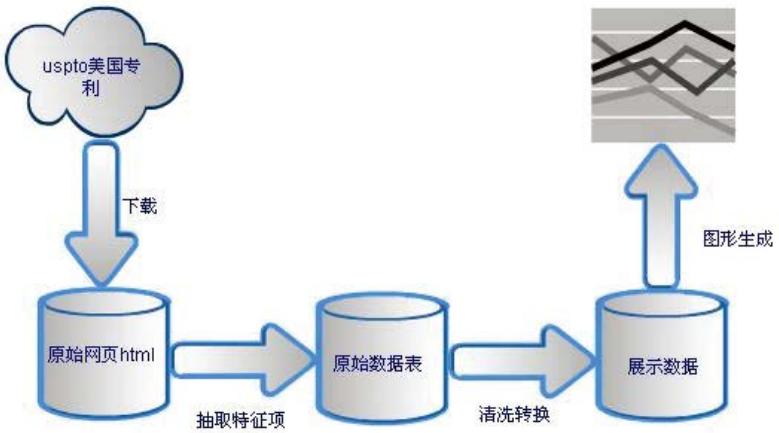


图2 数据处理流程图

### 3.2 系统实现

(1) 数据获取

美国专利数据库有比较悠久和完整的专利引文信息资源。本文就以美国专利数据库为数据源，检索“新能源汽车”领域的专利数据及引文信息。通过爬虫工具把符合检索条件的专利网页Html下载到本地数据库。然后用正则表达式来抽取专利特征项，如专利号、标题、文摘、发明人、专利权人、国际专利分类号（IPC）等信息，存入原始数据表。

(2) 清洗转换

为了提高数据的准确性，原始数据需要经过清洗才可使用，如把同一机构的不同表达方式进行统

一合并，去除噪音数据等。根据分析的需要，如果需要分析企业集团之间的引用关系等，就需要把集团下的子公司统一用集团名字进行标注。有些复合字段需要进行拆分，如发明人、IPC等。最后，按表1和表2的表结构样式进行存储。

表1 结点信息表

| 字段名 | 字段属性 | 字段长度 |
|---|---|---|
| ID | 整型 | 11 |
| 专利号 | 字符串 | 20 |
| 日期 | 日期型 | / |
| 标题 | 字符串 | 200 |
| 发明人 | 字符串 | 200 |
| 集团 | 字符串 | 100 |
| 专利权人 | 字符串 | 200 |
| 国际专利分类号 | 字符串 | 100 |
| 被引次数 | 整型 | 11 |

表2 结点关系表

| 字段名 | 字段属性 | 字段长度 |
|---|---|---|
| ID | 整型 | 11 |
| 父ID | 整型 | 11 |
| 子ID | 整型 | 11 |
| 引用次数 | 整型 | 11 |

表1存储树结点信息，比如研究专利之间的引用关系。表1中的字段存储了该专利的标题、日期、发明人、专利权人、IPC等信息，其中被引次数表示该专利被引用的次数，在图形上可以用结点的大小来表示。表二中存储结点之间的引用信息，其中“父ID”和“子ID”字段代表树中的父结点和子结点，都为表1中的ID字段值，父结点与子结点是一对多关系，每一个子结点只有一个父结点。“引用次数”字段用于集团之间的引证树，表示一个集团引证另一个集团次数的多少，在引证树上用结点间连线粗细来表示。

(3) 生成图形

把处理后的数据转换成图形显示出来是本系统核心模块，从以下程序片段可以大体看出系统生成图像的过程。具体步骤如下：

① 连接数据库。Prefuse支持从数据库中获取数据或从xml文件中读取数据等多种方式。

② 生成抽象的树形数据结构模型。

③建立可视化对象平台，随后的加载数据和进行参数设置都在此基础上进行。

④加载数据。

⑤设置样式，可通过对样式工厂的详细设计设置一些基本的参数，如结点大小、线粗细，有向或无向等。

⑥添加颜色描述。可根据数据库中某一字段的类别属性随机生成不同的颜色的结点，以区别不同类别之间的引用关系。

⑦ 添加一些动画效果，可更人性化、更友好地对树形结构图进行展示盒隐藏。

⑧最后建立展示对象，在Jframe对象上展示出来。

```
…
//连接数据库
datasrc = ConnectionFactory.
getDatabaseConnection(driverNam
e,dbURL, userName, userPwd);
//生成树数据
Table nodes = datasrc.
getData("select * from node");
Table edges = datasrc.
getData("select * from edge");
t = new Tree(nodes, edges,
"ID", "父ID", "子ID");
//建立可视化对象
Visualization m_vis=new
Visualization();
//加载数据
m_vis.add(tree, t);
//设置样式
m_vis.setRendererFactory(rf);
//添加颜色
m_vis.putAction("textColor",
textColor);
//添加动画
m_vis.putAction("animate",
animate);
//建立Display展示
Display display = new
Display(vis);
JFrame frame = new
JFrame("prefuse example");
frame.add(display);
…
```

## 3.3 系统功能

(1)界面介绍

引证树分析工具界面如图3所示。软件上边是菜单栏，主要包括“文件”、“数据”、“设置”和“帮助”四个菜单；中间为图形显示区；右侧从上至下依次是参数设置、全图预览和相关信息查看。

系统支持从“文件”菜单打开外部xml数据进行展示，同时也可以通过选择数据菜单来展示系统内置的新能源汽车领域的专利数据。通过参数设置可以调整图形显示的长短、粗细、间隔等属性。当引证

树较大时，可以通过预览区域看到引证树的全貌。把鼠标放在专利结点上，可以看到该专利的详细信息在“相关信息”区域显示出来。

(2) 主要功能

● 输入检索词，可以查看以此检索词为结点的专利引证树，左侧为被引，右侧为施引。

● 可以查看专利号之间的引证树或专利所属集团公司之间的引证关系。

● 引证图可以缩放大小、拖动，可以通过点击结点，逐级打开下层的引证关系，可以拖动结点，可以调整树的指向方向等一系列方便的查看技巧。

● 可以调整图形参数，以达到美观的效果。

● 可以查看全图预览，把握整体结构。

● 可以查看结点详细信息。

● 若是集团公司之间的引证树，结点大小代表被引次数的多少，线的粗细表示施引者引用被引者专利次数的多少。

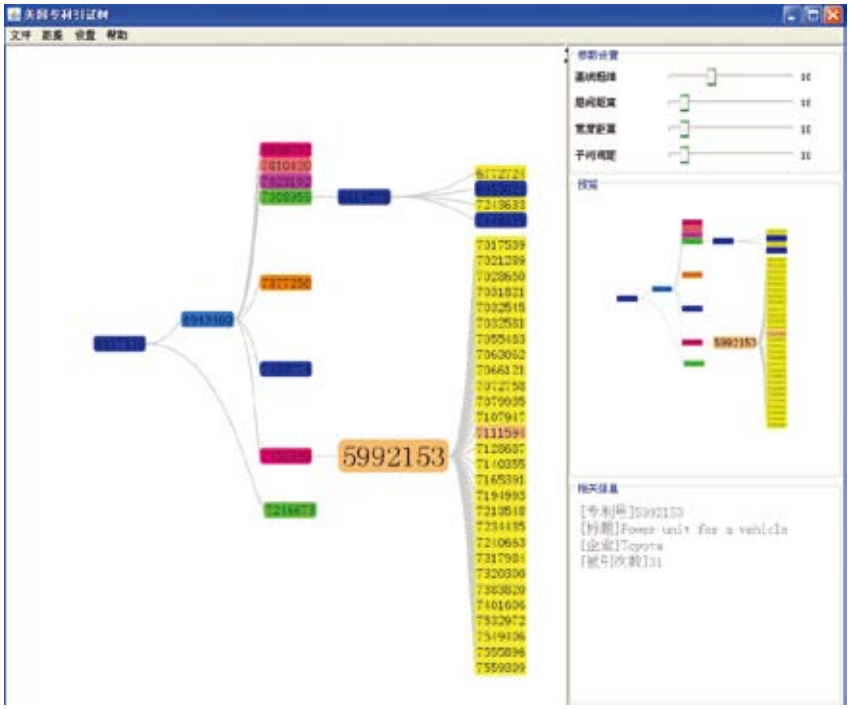


图3 引证分析工具界面

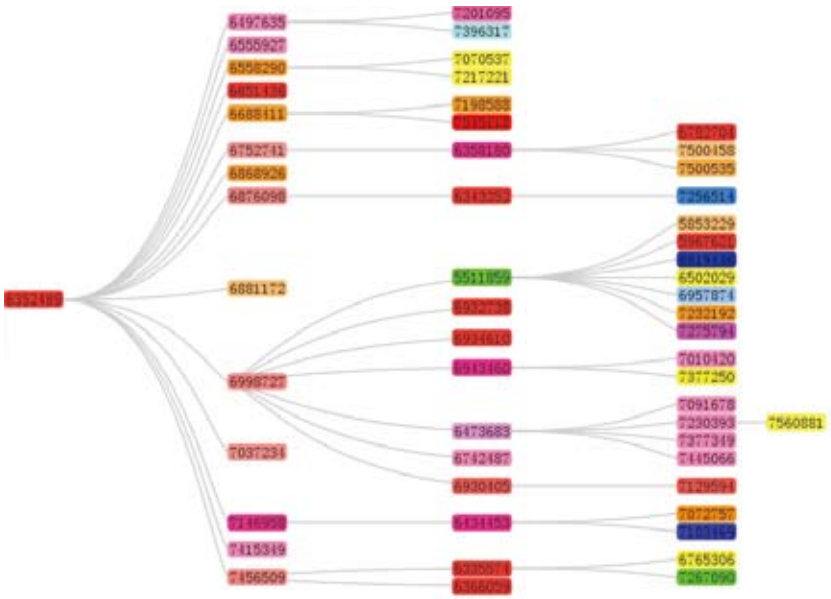

图4 专利引证树一

## 4 引证树在新能源汽车技术领域的实证分析

本文以新能源汽车领域专利文献为研究对象，检索出与新能源汽车相关的2000年至2009年十年的专利及其引文数据，经过数据清洗转换后存入数据库。主要包括Ford、Honda、Toyota、General、Nissan、Panasonic等6家公司的数据，通过引证树来分析这几家在新能源汽车领域的竞争态势和专利布局策略。

(1) 查看被引情况。通过检索某条专利，如检索专利号为“6352489”的美国专利，可以查看以此专利为根结点的专利树引证情况，鼠标放在子结点上，在“相关信息”区会看到该结点专利的详细信息，由此发现该专利的被引过程和技术发展路线，不同颜色的结点，代表不同的公司，如图4所示。

(2) 发现竞争对手专利布局。如图5，蓝色的是Honda公司专利，橙色是Ford公司专利，通过引证树可以发现Honda公司的专利“6866649”被大量的Ford公司专利引用，存在专利围剿现象，由此图

可见几家公司在该子领域的争夺十分激烈。

(3) 查看公司的专利布局是否完善。如图6所示，黄色的专利是GM公司的，在专利“6488345”后，该公司又进行了大量的围绕该专利的专利申请，正确的专利布局使得公司能更好地占领技术制高点，加大竞争对手的进入障碍。

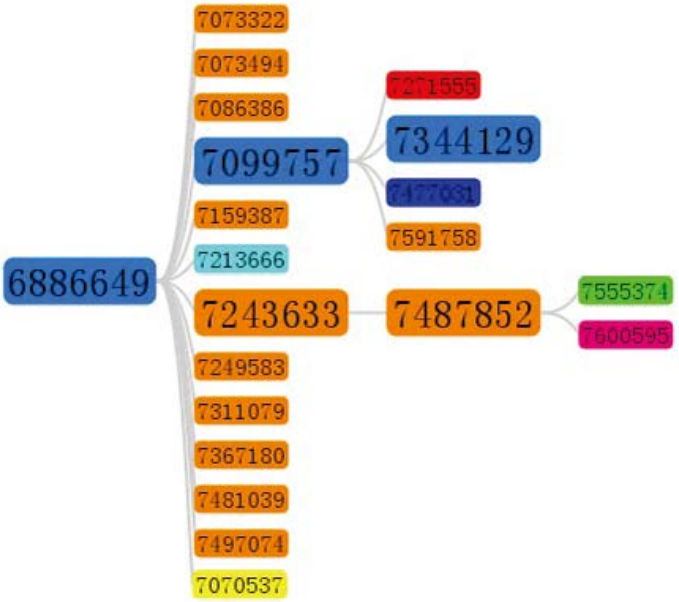


图5 专利引证树二

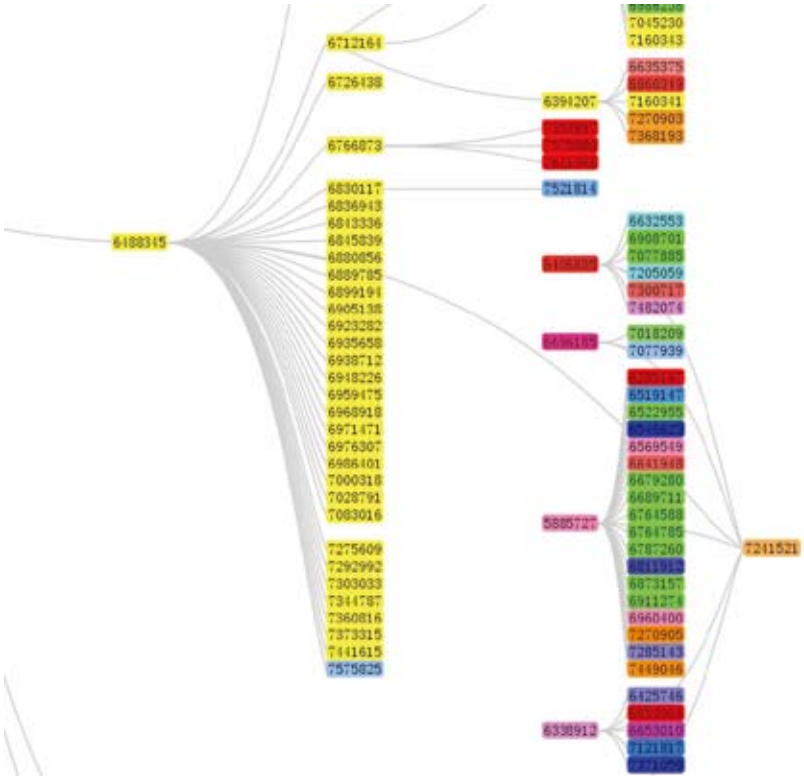


图6 专利引证树三

(4) 通过专利引证树可以发现核心专利。通过引证树可以一目了然地看到该领域的核心专利都有哪些，及高被引专利布局，从而了解该领域的技术分布状况。如图7，通过对某子领域的引证树分析，可以迅速发现该领域的核心专利有“6330498”、“6359404”、“6190282”、“6278951”等。然后可以对这些核心专利进行进一步的分析。

## 5 结语

本文运用信息可视化的方法，基于美国专利引文数据，描述了如何设计和开发出一个专利引证树的过程，并进行了基本的应用分析。通过专利引证分析，分析人员可以在大量的专利信息中迅速把握最相关的文献信息，建立专利之间的关联关系，掌握技术发展的脉络。通过对某公司专利的分析，也可以了解该公司的核心专利保护策略，进而得到竞争对手在其外围的专利布局以及核心专利技术的最新发展方向，为企业规避竞争对手的核心技术、企业技术引进、制定研发路线提供决策依据，为研究人员提供突破已有专利技术的保护限制，寻求新的研发机会提供决策支持[6]。

专利引证分析正在越来越多地被人们使用，逐步成为专利计量分析的核心内容。但由于一些客观原因，国内对专利分析工具的开发与国外还有较大的差距。比如，很多从事情报分析的人员技术基础比较薄弱，而技术开发较强的人员又缺乏基本的情报学分析理论和思想。

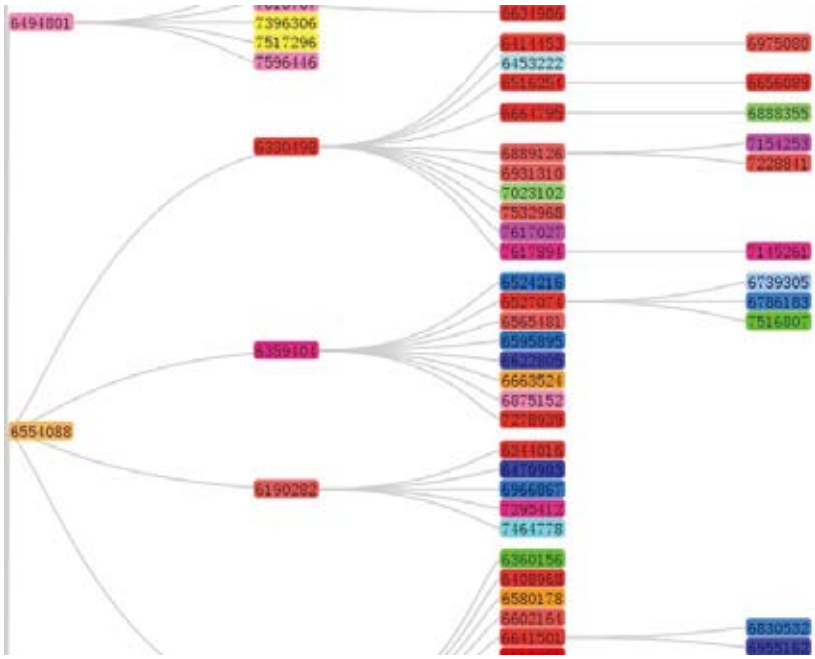


图7 专利引证树四

还有，国外现有的专利分析工具由于价格昂贵等原因也导致国内使用普及率较低，阻碍研究人员对专利文献的深度挖掘利用。本文正是针对这一课题提出了一种可行的整体解决方案，叙述了从专利数据的获取、转换、可视化展示的全过程，并详细介绍了一种用于可视化开发的开源工具包Prefuse。这些分析工具和过程同样也可以扩展应用在其他领域，如文献之间的引用分析等。

**参考文献**

[1] 王庆稳,邓小昭. 专利引文分析及其应用研究[J]. 现代情报,2008(4):189-192.
[2] 任智军,朱东华,谢菲. 专利引用分析方法研究[J]. 商业时代,2007(14):102-104.
[3] 周宁,张玉锋,张季义. 信息可视化与知识检索[M]. 北京:科学出版社,2005.
[4] prefuse[EB/OL]. [2010-7-21]. http://prefuse.org/.
[5] prefuse | documentation > manual > introduction > toolkit structure [EB/OL]. [2010-7-21]. http://prefuse.org/doc/manual/introduction/structure/.
[6] 吴正. 可视化工具在专利分析中的应用[J]. 数字图书馆论坛,2009(10):60-67.

**作者简介**

张兆锋(1979- )，硕士，工程师。研究方向：信息系统和信息可视化。通讯地址：北京市复兴路15号 中国科学技术信息研究所 信息技术支持中心 100038。E-mail: zhangzf@istic.ac.cn
桂婕(1976- )，博士，助理研究员。研究方向：专利分析和科技创新管理。通讯地址同上。E-mail: guij@istic.ac.cn
乔晓东(1965- )，硕士，研究员，研究方向：信息服务和信息资源管理。通讯地址同上。E-mail: qiaox@istic.ac.cn
朱礼军(1973- )，博士，副研究员，研究方向：知识组织。通讯地址同上。E-mail: zhulj@istic.ac.cn
李鹏(1979- )，硕士，助理研究员。研究方向：智能信息处理。通讯地址同上。E-mail: lipeng_cn@istic.ac.cn

## Design and Implementation of a Patent Citation Analysis Tool

Zhang Zhaofeng, Gui Jie, Qiao Xiaodong / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Patent Citation Analysis is more and more important in patent analysis. This paper introduces the design and implementation of a tool for patent citation analysis, including data acquisition, data cleaning, data transformation and generating graphics. And, an open source visualization toolkit is also introduced, which will be helpful for researchers engaged in literature research through visualization method.

Keywords: Patent citation, Information visualization, Data mining, Prefuse

（收稿日期：2010-08-15）