

## 文本信息可视化模型研究<sup>1)</sup>

周宁<sup>1</sup> 张会平<sup>1</sup> 金大卫<sup>1,2</sup>

(1. 武汉大学信息资源中心, 武汉 430072; 2. 中南财经政法大学信息学院, 武汉 430064)

**摘要** 本文针对文本信息资源的特征, 提出了一个基于 XML 的文本信息可视化的通用模型, 详细介绍了模型的三个对象空间——XML 文档库、XML 特征库和可视化对象以及三项关键技术——中文分词、文本分割和可视化映射, 并结合实例验证了模型的实用性、易扩展性以及可移植性。

**关键词** 文本 信息可视化 中文分词 文本分割

### Research on Text Information Visualization Model

Zhou Ning<sup>1</sup>, Zhang Huiping<sup>1</sup> and Jin Dawei<sup>1,2</sup>

(1. Research Center of Information Resources, Wuhan University, Wuhan 430072;

2. Information school, Zhongnan University of Economics and Law, Wuhan 430064)

**Abstract** According to characteristics of text information resources, this paper brings forward a common model of text information visualization. We introduce its three object spaces: XML document library, XML characteristic library and visual object and three key technologies: Chinese segmentation, text segmentation and visualization mapping. In the end, we testify its practicality, expansibility and transplant with an example.

**Keywords** text, information visualization, chinese segmentation, text segmentation

20 世纪 90 年代, 信息可视化问题成了国际热点研究课题之一, 信息检索、超文本、WWW、数字图书馆以及人机接口等领域都在关注这一课题<sup>[1]</sup>。信息可视化不仅给信息以形象, 为我们提供直观的结果, 方便我们观察和记忆; 而且能够揭示信息之间的关联, 有利于我们形成整体概念、发现隐含的知识<sup>[1-3]</sup>。

信息可视化的基本过程是: ①数据准备, 即确定和获取可视化的原始数据, 形成原始数据空间; ②数据析取, 即从原始数据中析取需要可视化的数据, 形成可视化数据空间; ③可视化映射, 即采用一定的映射算法把可视化数据空间映射到可视化对象<sup>[1-4]</sup>。信息资源按其媒体形式可以分为三类: 文本信息、音

频信息和视频信息<sup>[5]</sup>。不同类型的信息资源, 其描述方式和特征并不相同, 因此, 形成的可视化数据空间和选取的可视化对象也不相同。本文针对文本信息资源的特征, 提出了文本信息可视化的通用模型, 并结合实例验证了这一模型的实用性、易扩展性以及可移植性。

## 1 引言

文本信息资源大量存在于全文数据库、电子期刊、Web 数据库、WWW 页面、数字图书馆当中。文本信息资源具有以下特点: ①内容多样, 例如, 图书、报刊、科技报告、会议论文、学位论文、专利说明书、

收稿日期: 2006 年 3 月 22 日

作者简介: 周宁, 男, 1943 年生, 教授, 博士生导师; 张会平, 男, 1982 年生, 博士研究生; 金大卫, 男, 1978 年生, 博士研究生。

1) 教育部哲学社会科学重大课题攻关项目(05JZD00024)和国家自然科学基金项目(70473068)资助。

政府报告等;②格式繁杂,例如,TXT、HTML、DOC、PDF、CAJ、VIP等;③数量巨大,例如,我们以“可视化”为标题关键字查询中国学术期刊全文数据库时,系统返回3221条记录(2006年3月5日);④应用广泛,我们时时刻刻在使用不同的文本信息。

文本信息资源具有外部特征和内部特征:外部特征包括资源的创建者、创建时间、来源等;内部特征包括标题、分类、关键词、摘要、全文等<sup>[5,6]</sup>。因此,在描述文本信息时需要同时考虑其内部和外部特征。

## 2 文本信息可视化模型

在充分考虑文本信息资源特征的基础之上,我们提出了一个基于XML的文本信息可视化模型,如图1所示。XML相关标准不仅集显示与内容于一体,而且是一项数据交换的标准。因此,我们提出的基于XML的模型具有强移植性和扩展性。从体系结构上看,该模型类似于网络协议体系的沙漏模型,具有较强向下兼容、向上扩充的能力。不同格式的文本信息如TXT、HTML、PDF通过格式转换、抽取、识别得到规范化的XML文档,进而生成XML文档库。XML文档通过特征抽取(包括中文分词、文本分割以及词频统计)得到相应的XML特征文档,进而生成XML特征库。XML特征库通过不同的可视化映射算法生成不同的可视化对象,方便我们检索和利用文本信息。本节我们首先讨论三个对象模型——XML文档库、XML特征库以及可视化对象,下节我们将对模型中涉及关键技术做详细讨论。

### 2.1 XML文档库

通过对原始文本进行转换生成的XML文档库的格式,我们使用XML Schema进行定义和规范,如图2所示。“passages”元素由一系列“passage”子元素组成,代表相应的文档集合。而对每个具体的文档,我们使用属性“id”、“title”、“subject”、“author”、“source”以及“date”分别描述了文档的“惟一标识”、“标题”、“主题”、“作者”、“来源”以及“创建时间”,使用三个子元素“keyword”、“abstract”以及“context”描述文档的“关键词”、“摘要”以及“全文”。

### 2.2 XML特征库

我们同样使用XML Schema定义和规范了XML特征库的文档格式,如图3所示。特征库与文档库

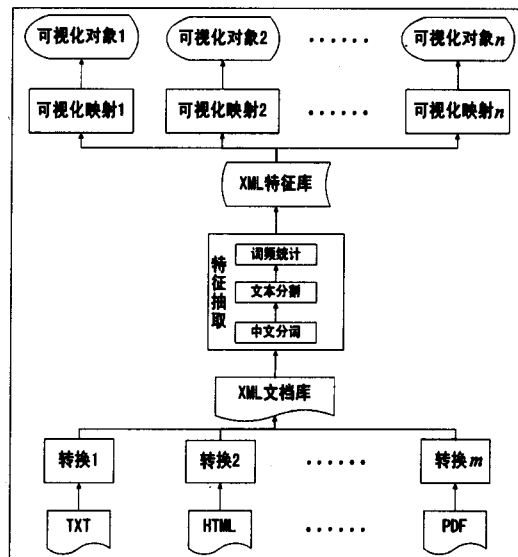


图1 文本信息可视化模型

```
<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="passages">
<xs:complexType>
<xs:sequence>
<xs:element name="passage" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
<xs:element name="keyword" type="xs:string" />
<xs:element name="abstract" type="xs:string" />
<xs:element name="context" type="xs:string" />
</xs:sequence>
<xs:attribute name="id" type="xs:string" />
<xs:attribute name="title" type="xs:string" />
<xs:attribute name="subject" type="xs:string" />
<xs:attribute name="author" type="xs:string" />
<xs:attribute name="source" type="xs:string" />
<xs:attribute name="date" type="xs:date" />
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>
```

图2 XML文档库XML Schema

的区别在于,我们不再使用“关键词”、“摘要”以及“全文”来描述文档,而是使用一系列词在文档各个部分的词频统计信息来描述文档。词的类型包括三类:标题词、关键词以及自由词。标题词通过对文档标题分词得到;关键词直接来自原来的关键词列表;自由词通过对全文分词得到。对于规范、完整的原始文档具备关键词列表,但也存在大量文本信息并没有给出关键词列表,有些甚至没有标题。另外,部分文档标出的关键词并没有在文档中出现。因此,为了准确地描述文档信息,我们引入了自由词作为有力补充。同时,我们不仅统计词在整篇文档中出现的频率,而且通过对文档分割划分逻辑段落、形成

文档的子主题列表,统计词在各个子主题中的频率分布情况。这样,我们对文本的描述直接进入文档的内部,在用户查询时能够定位到文档的内部,使其更为快捷地了解 and 把握文档内容结构,找到其所需部分直接阅览。

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="passages">
<xs:complexType>
<xs:sequence>
<xs:element name="passage" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
<xs:element name="wordsFreq">
<xs:complexType>
<xs:sequence>
<xs:element name="wordFreq" maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
<xs:element name="partFreq" maxOccurs="unbounded">
<xs:complexType>
<xs:attribute name="part" type="xs:int"/>
<xs:attribute name="freq" type="xs:int"/>
</xs:complexType>
</xs:element>
</xs:sequence>
<xs:attribute name="word" type="xs:string"/>
<xs:attribute name="totalFreq" type="xs:int"/>
<xs:attribute name="wordType" type="xs:string"/>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:sequence>
<xs:attribute name="id" type="xs:string"/>
<xs:attribute name="title" type="xs:string"/>
<xs:attribute name="subject" type="xs:string"/>
<xs:attribute name="author" type="xs:string"/>
<xs:attribute name="source" type="xs:string"/>
<xs:attribute name="date" type="xs:date"/>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>
```

图3 XML 特征库 XML Schema

## 2.3 可视化对象

可视化对象是指从 XML 特征库中的数据对象生成的由视觉属性构成的对象。视觉属性包括位置、形状、方向、色彩、纹理、灰度等级、尺寸等<sup>[7,8]</sup>。我们可以在特征层、文档层、文档集合三个层次构建可视化对象。在特征层,我们可以为不同的主题选择不同的颜色或图符以在视觉上区别不同主题的文档,我们可以使用不同的灰度等级代表不同的日期等。在文档层,我们可以用树形结构显示一篇文档,包括三级结构:文档节点,隐喻文档情况;词整体统计节点,隐喻一个词在整篇文档中分布的情况;词部分统计节点,隐喻词在各个部分分布的情况。在文档集合层,我们可以把某一作者的所有文档作为一个集合,进而使用树或者网状结构对其进行可视化

呈现。这里只是举例说明,具体采用什么可视化对象,可以视具体情况而定。

## 3 关键技术

在我们提出的模型当中涉及到三项关键技术——中文分词、文本分割和可视化映射,本节我们将详细地探讨它们的发展情况及其在模型中的应用方案。

### 3.1 中文分词

中文与西文不同,中文的词语之间既无特殊的空格也无特殊的间隔标志,也就是说词之间没有明显的形态界限。几十年来,关于中文分词的研究一直长盛不衰。中文分词大致可以分为三类:基于词典的分词方法,基于统计的分词方法以及混合的分词方法<sup>[9]</sup>。

基于词典的分词方法的三个要素是分词词典、文本扫描顺序和匹配原则<sup>[10]</sup>。文本的扫描顺序有正向扫描、逆向扫描以及双向扫描。匹配的原则有最大匹配、最小匹配、逐词匹配以及最佳匹配。基于词典的分词方法的优点是易于实现。其缺点是匹配速度慢、存在歧义切分问题。而且词典中词目的选择和词目的数量直接影响分词的结果。

基于统计的分词方法的主要思想是:词是稳定的字的组合,因此在上下文中,相邻的字同时出现的次数越多,就越有可能构成一个词<sup>[11]</sup>。因此字与字相邻共现的频率或概率能够很好反映成词的可信度。应用的主要统计量或统计模型有:互信息、N元文法模型、神经网络模型、隐 Markov 模型、EM 模型、关联词统计语言模型以及最大熵模型等。基于统计的分词方法的优点是不受待处理文本领域的限制、不需要词典、有效解决歧义问题。其缺点是需要大量的训练文本来建立模型的参数、计算量大。而且训练文本的选择直接影响分词的效果。

混合的分词方法就是综合利用多种方法来提高分词的精度。例如,文章<sup>[12]</sup>提出的一种提高中文分词精度的多步处理策略,包括:消除伪歧义、部分确定性切分、数词串处理、重叠词处理、基于统计的未登录词识别以及使用词性信息消除切分歧义的一体化处理。

目前的中文分词工具一般都是采用了混合分词方法,不仅维护了一个词典,而且采用多种统计模型消除切分歧义以及识别未登录词。因此,我们给出

的模型选择采用混合分词方法。同时,由于特征库里存储了词条的词频统计信息,可以使用这些信息来更新词典,把其中频率较高而未在词典中的词条加入到词典当中,实现词典的自动更新。

### 3.2 文本分割

文本分割是指对一篇文本的各个段落按照语义关系进行分割,将各个自然段落进行归并,使得文章中大意属于一个子主题的段落归并到一个语义段落,这样一篇文本就分割成若干语义段落。在一些文本中,特别是说明文和议论文,会有一些小标题,由于存在这些物理标记,对于这种类型的文本进行分割(称为显式文本分割)比较容易,直接按作者的标识即可。但是,大多数文本中并没有这些明显的物理标记,对于这些文本的分割(称为隐式文本分割)则较为复杂。

当前,针对隐式文本分割人们提出了不同的模型。例如,Hearst 提出的 TextTiling 算法<sup>[13]</sup>、Reynar 的 Dotplotting 模型<sup>[14]</sup>、Chen 的融合模型<sup>[15]</sup>、Beeferman 的统计语言模型<sup>[16]</sup>以及 Blei & Moneo 的隐 Markov 模型<sup>[17]</sup>。前两种方法主要是基于不同段落中相同词语的数目以及词语密度来进行文本分割。Chen 的模型集成了短语复用、语义重复以及词语的 tf-idf 等特征来进行段落之间相似性的判断,由于融合了多种特征,这一模型能够进一步改善文本分割的准确度。Beeferman 则利用了统计语言模型,通过与领域相关的线索词的结合,这一模型表现出了很好的主题跳转判别性能。将文本段落之间的关系映射成一个标记序列,隐 Markov 模型对标记序列具有较好的分割效果。

文本分割的关键问题是确定子主题的跳转。往往子主题之间既存在着区别性,也存在着相似性。我们根据区别程度的大小将两个相邻的段落分开,但这个阈值很难确定。另外,相似度的计算也是决定切分位置的关键因素,不同的相似度计算方法必然导致不同的切分位置,因此还需考虑相似度的计算方法。

在我们的模型当中,对于存在物理标记的文本直接按照作者的标识进行文本分割。而对于没有明显物理标记的文本,由于整个模型以词频统计为基础因而可以采用 TextTiling 算法。同时,可以利用词条的全局统计信息对 TextTiling 算法进行优化:一是,考虑利用词条在整篇文档的词频信息;二是由于在特征库里存储了词条的统计信息,因此还可以考

虑利用词条在整个特征库里的统计信息。例如,我们可以使用这些统计信息调整在文档分割时词条的权值,对于词频较高的词条给以较高的权值,相应地,对于词频较低的词条给以较低的权值。

### 3.3 可视化映射

最早由 Shneiderman 教授概述的按照数据类型把信息可视化技术方法分为以下七类:一维数据可视化、二维数据可视化、三维数据可视化、多维数据可视化、时间数据可视化、层次数据可视化和网状数据可视化<sup>[18]</sup>。其中,文本是简单的线性数据,属于一维数据可视化的范畴。但是,在我们提出的模型当中,通过对原始文档特征抽取后形成的特征库则是多维数据并且具有时间性、层次性和网络性。我们不仅描述了文本的外部特征,还从深层次揭示了其内部特征,从多个角度、不同层面形成了多维数据空间。由于刻画了文档的时间属性,对于对时间敏感的文本,如新闻,我们可以用时间序列映射算法形成可视化对象。我们对文档内部特征的描述本身就是基于层次结构的,因此具有层次性。同时,我们可以根据组合多个层次结构形成网络,例如,我们可以把不同作者的不同文档进行组合,形成网状结构,通过网状数据可视化技术发现其中的隐含结构。因此,在我们的模型当中,可以使用目前所有的较为成熟的可视化映射算法来对特征库进行可视化映射,这就说明我们给出的模型满足通用性要求,并具有较强的扩展性和兼容性。

例如,我们可以使用图符标识法描述文本的主题内容,使用不同的图符清晰、直观地表达文档的主题,让用户一目了然。我们可以使用高维空间描述法映射文档的词条分布,方便用户把握文档的内容结构。在下一节中,我们使用了层次信息可视化工具 Treemaps 在文档和文档集合两个层次实现了可视化映射,详细见下一节内容。另外,我们也可以使用其他层次数据可视化技术如双曲树、节点连线式树、三维形式的树结构等,以及网状数据可视化映射算法实现对特征库的可视化映射。

## 4 实例分析

最后,我们通过具体的例子来验证所提出模型的实用性、易扩展性以及可移植性。我们从中国学术期刊全文数据库中获取 10 篇关于“信息可视化”的文章,如表 1 所示。通过对这些文档格式转换、特

征抽取得到满足上述特征库模式的 XML 文档。我们首先利用图符标识法、高维空间描述法对其进行可视化呈现。这两种方法充分反映了我们所提出模型的实用性<sup>[5,6]</sup>。为了进一步说明模型的实用性、易扩展性以及可移植性,我们把得到的 XML 特征文档导入 Treemaps 工具当中做进一步的验证。

表 1 文章列表

序号	标题	来源	年(期)
1	信息可视化系统的 RDV 模型	情报学报	2004(5)
2	信息可视化的基本过程与主要研究领域	情报科学	2004(1)
3	信息可视化在信息管理中的新进展	现代图书情报技术	2003(4)
4	数字图书馆信息可视化的研究框架	沈阳教育学院学报	2005(3)
5	文献信息可视化研究	情报学报	2003(4)
6	信息可视化——知识服务网站的新形象	情报理论与实践	2005(6)
7	信息提供的可视化研究	情报科学	2004(3)
8	信息资源描述与存储的可视化研究	情报科学	2004(1)
9	同引分析与可视化技术	情报科学	2005(4)
10	引文分析可视化研究	情报杂志	2004(11)

Treemaps 是由美国马里兰大学计算机学院人机接口实验室的 Shneiderman 教授提出的基于空间填充算法的层次信息可视化工具<sup>[19]</sup>。Treemaps 使用嵌套的矩形来呈现层次关系,使用矩形的面积大小和填充颜色来隐喻不同的变量。经过一系列优化和改进,Treemaps 已经得到了广泛应用。该实验室开发的 Treemaps 工具 Treemap 自版本 4.0 开始支持可扩展的层次结构进一步增强了 Treemaps 的实用性<sup>[20]</sup>。我们从该实验室的网站<sup>[21]</sup>上下载了其最新版本 Treemap4.1 来验证我们的模型。

Treemap4.1 支持 XML 格式,但是其定义的模式与我们的特征库模式并不一致,因此,我们首先要进行 XML 模式转换。XML 模式之间的转换是相当容易的,也说明我们提出的基于 XML 的模型具有较强的可扩展性以及可移植性。

Treemap4.1 功能相当强大,在这里我们仅给出两个例子。首先,我们使用它来对 10 篇文档分类,我们定义两个层次结构:第一层“年份”、第二层“来源”,并且使用矩形颜色的深浅隐喻不同的年份,矩形的大小不作隐喻设置,结果如图 4 所示。从图 4 我们可以看到文档之间的关系非常清晰、直观,与表 1 相比,更加易于观察。我们能够一眼看到 10 篇文章在年份、来源上的分布情况。例如,2004 年有 5 篇,其中《情报科学》3 篇。

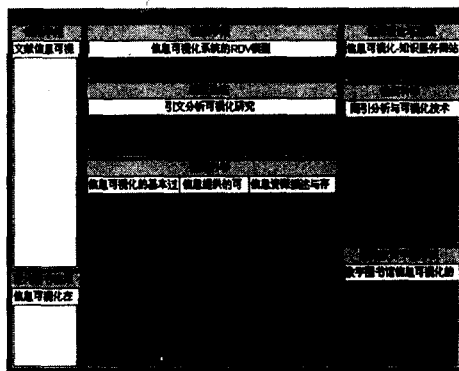


图 4 Treemap4.1 可视化呈现文档分类

第二个例子,我们来说明可视化相关词在 10 篇文档中各个部分的分布情况。这里,我们定义的层次结构是:第一层“词”、第二层“标题”、第三层“部分”,并且使用矩形颜色深浅隐喻不同的年份,使用矩形的大小隐喻词频的数量。通过上述定义,我们会得到所有词在 10 篇文档中的各个部分分布情况,为了清晰看到效果,我们双击隐喻“信息可视化”的矩形,Treemap4.1 就会放大(zoom in),屏幕里只显示该词的分布情况,如图 5 所示。从图 5 我们知道“信息可视化”在文档《信息可视化的基本过程与主要研究

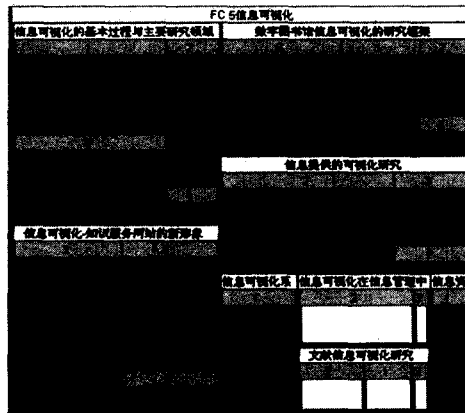


图 5 Treemap4.1 可视化呈现词频统计

究领域》中出现的频率最高,而且主要分布在1,3两个部分。对照原文,我们发现这两个部分篇幅最长,而且是讨论信息可视化的基本问题——概念和研究范围。如果我们关注可视化模型的研究,我们可以放大“模型”这个词的分布情况,就可以找到相关文档的相关部分。这就说明我们的模型具有很强的实用性。

## 5 结束语

在充分考虑文本信息资源特征的基础之上,我们提出了一个基于XML的文本信息可视化模型,并使用图符标识法、高维空间描述法以及Treemaps工具验证了这一模型的实用性、易扩展性以及可移植性。我们后续的研究将以此模型为基础,致力于开发更为有效、适合网络应用的可视化映射方法,利用可视化的方法解决人们在文本信息资源利用中的问题。

## 参 考 文 献

- [1] Chaomei Chen. Information Visualization beyond the Horizon-Second Edition. London:Springer-Verlag,2004.
- [2] 宋绍城,毕强. 信息可视化的基本过程与主要研究领域. 情报科学,2004,22(1):13-18.
- [3] 周宁,杨峰. 信息可视化系统的RDV模型研究. 情报学报,2004,23(5):619-624.
- [4] 文燕平. WWW信息检索可视化原理研究. 现代图书情报技术,2005(4):10-13,50.
- [5] 周宁. 信息资源描述与存储的可视化研究. 情报科学,2004,22(1):9-12,18.
- [6] 周宁,文燕平,刘玮. 文献信息可视化研究. 情报学报,2003,22(4):468-471.
- [7] Colin Ware. Information Visualization: Perception for Design. San Francisco: Morgan Kaufmann,2000.
- [8] Card S K, Mackinlay J. The Structure of the Information Visualization Design Space. Proceedings of the 1997 IEEE Symposium on Information Visualization:92.
- [9] 付国宏,王晓龙. 汉语词语边界自动划分的模型与算法. 计算机研究与发展,1999,36(9):1143-1147.
- [10] 张春霞,郝天永. 汉语自动分词的研究现状与困难. 系统仿真学报,2005,17(1):138-143,147.
- [11] 曹倩,丁艳,王超,等. 汉语自动分词研究及其在信息检索中的应用. 计算机应用研究,2004(5):71-74,91.
- [12] 赵铁军,吕雅娟,于浩,等. 提高汉语自动分词精度的多步处理策略. 中文信息时报,2001,15(1):13-18.
- [13] Hearst M A. Segmenting Text into Multi-paragraph Subtopic Passages. Computational Linguistics, 1997,23(1):33-64.
- [14] Reynar J C. An Automatic Method of Finding Topic Boundaries. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, 1994:331-333.
- [15] Chen Qingcai, Wang Xiaolong, Liu Bingquan. Subtopic Segmentation of Chinese Document: An Adapted Dotplot Approach, ICMC'02,2002,3:1571-1576.
- [16] Beeferman D, Berger A, Lafferty J. Statistical Models for Text Segmentation. Machine Learning, 1999, 34(1-3):177-210.
- [17] David M B, Pedro J Moreno. Topic Segmentation with an Aspect Hidden Markov Model. Research and Development in Information Retrieval, 2001:343-348.
- [18] Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. IEEE Symposium on Visual Language, 1996:336-343.
- [19] Johnson, Brian, Shneiderman, Ben. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In Proceedings of the 1991 IEEE Information Visualization: 284-291.
- [20] Chintalapani G, Plaisant C, Shneiderman B. Extending the utility of treemaps with flexible hierarchy. In Proceedings of the 2004 IEEE Information Visualization: 335-344.
- [21] <http://www.cs.umd.edu/hcil/treemap/> (Accessed Dec. 20, 2005).

(责任编辑 王建平)