课程设计报告

选题名称:

基于文本可视化分析的

研究领域主题地图

学 院：\_\_\_\_\_软件学院\_\_\_\_\_\_\_\_\_\_\_\_

专 业：\_\_\_\_\_软件工程\_\_\_\_\_\_\_\_\_\_\_\_

小组人员：

姓名：\_\_高梦雅\_\_学号：\_\_3013218105\_工作量：\_\_33.3\_\_\_%

姓名：\_王善宇\_\_\_学号：\_3013218120\_\_工作量：\_\_33.4\_\_\_%

姓名：\_刘林\_\_学号：\_\_\_3013218111\_\_\_工作量：\_\_33.3\_\_\_%

天津大学软件学院

2015年12月29日

## 设计目的及意义：

对科学知识体系结构及其演化规律的研究一直是科技界发现研究热点、遴选优先领域、规划学科发展和制定科学政策的重要基础。本次项目旨在通过对于几本期刊近十年的文本进行分析，提取每本期刊的主要研究领域，并据此绘制科学结构地图，以快速、全面和形象的把握科学总体态势，分析各个期刊在研究领域上的侧重点，以及发展趋势。揭示不同领域期刊在主题和主体之间的关系，以及不同的研究领域之间的内在联系。

本次项目主要运用文献计量学的理论和方法绘制结构图，并运用关键字提取、识别等技术分析不同期刊的重点研究领域，探究不同研究领域的学科交叉性。

我们选取了TVCG、TMM、TKDE这三本IEEE的期刊，这三本期刊分别代表了三个不同的研究领域，即可视化研究，图形处理，数据挖掘。通过对于这三本期刊近十年所有文章的主题进行提取和分析，将相应的文本进行可视化处理，分析这三个不同研究领域中研究主题的交叉情况。同时根据时间的变化，在可视化方案中展现每本期刊研究领域主题的变化情况。

## 主要方法与技术：

1. 数据源

IEEE的三本期刊

TVCG : IEEE Transaction on Visualization and Computer Graphics

TKDE : IEEE Transaction on Knowledge and Data Engineering

TMM : IEEE Transaction on Multimedia

1. 数据处理

1). 获取方法：

* 1. 整体思路：在后台模拟HTTP的GET请求获取IEEE网站关于每个期刊的对应页面的html文本信息，对于每个期刊遍历访问2006～2015年期间10个Volume的每个Issue。通过使用Java的一个工具包Jsoup，将每次获取到的html页面进行解析，分析出相应的dom节点，截取相应节点中的文字内容。其中包括对应期刊的名称，发布日期，文章名称，文章对应的关键字，文章的引用文献信息。将其保存到定义好的实体工具类中，将数据全部解析完成后，将其转换成Json格式的数据，并保存到本地。
  2. 输入：通过url中的punumber参数即可定位每本期刊相应信息的地址，通过后台发送HTTP请求，请求时通过改变传入的punumber参数即可获取不同的期刊。
  3. 输出：包含期刊名称，年份，期数，文章名称，文章关键字，文章引用文献的json格式数据。

2). 清洗方法

1. 整合数据

从数据源获取的数据是单篇论文的基本信息，我们定义干净的数据为对在可视化过程中有实际意义的格式化数据。那么我们要得到的干净数据需要从论文中提取，需要对数据进行时间上的归类，杂志归属上的归类，研究领域上的归类，主要信息数据和次要信息书举得归类

1. 聚类

聚类主要采用的是直接关系的聚类和间接关系的聚类：

直接关系的聚类是指两篇论文存在相同的关键字的关系，间接关系是指两篇论文存在祖先继承关系的相同关键字，祖先继承关系来自于引用关系，也就是通过当前文章所引用的文章的关键字与另一篇论文匹配上的关系，但在实际上很多论文的引用格式不尽相同，甚至在同一篇论文当中的引用格式可能都有多种，所以在匹配的过程存在一定的误差。我们主要采取了以下措施来实现一个合理的模糊匹配。

*1.不同国家地区的人具有不同的拼写习惯，会导致同一单词中的‘s’和‘z’相互替换，所以诸如此类的字母我们会允许一个阈值范围内的错误而依旧认为是正确匹配。*

*2. 在匹配过程中标点诸如连字符以及空格在匹配中没有实际意义，所以直接去掉，不参与匹配过程*

*3. 对匹配的单词给予一定的权值，增加动词和名词的权值，降低一些常用的形容程度的形容词和助词以及介词的权值，因为后者往往没有实际意义，并且判断匹配成功的标准为是否达到一定的匹配率*

*然后根据匹配到的关系，一些直接间接用到一些关键词的文章化作 一个分类，作为一个集合，我们对于这种集合，规模达到一定常量的 集合作为一个领域。以上过程作为聚类过程。*

3). 分类

根据之前的聚类，我们对每个文章进行分类，其实在聚类过程中已经可以完成了分类操作。

4). 数据存储

采用JSON对清洗的数据进行结构化存储，利用Java中json-lib.jar的包对数据进行解析，利用Java对数据进行处理，导出到JSON如下“：



1. 可视化方法
   1. 采用D3.js开源库配合 javascript 等前端语言进行可视化方案的实现。
   2. 采用了pack图作为主体，采用了可交互式的可视化方案，能够通过放大缩小适应性满足用户对于数据详略解读的需求
   3. 采用poly-sunburst图，是一种借鉴了折线图和sequence-sunburst两种可视化视图实现的一种新的数据布局，用于体现不同领域之间的交叉性，能够支持单条折线的剖分，能够实现旋转，方便重点关注一些领域放于利用观察的视线范围
   4. 采用色彩映射，通过三原色的交融体现三个领域的交叉性，用颜色的交融清洗的体现学科间的交叉性，一目了然。利用色彩的深浅反应对于领域的关注度

## 三、设计步骤：

可视化设计的目的，是为了能够清楚的表示出各杂志的热点研究领域（以下称为热点领域），以及不同杂志之间的领域关联、相互参考（以下称为交叉领域）。因此，可视化的设计要做到以下几点：

1. 存在一个总体科学地图，能够同时表示出在一个时间区间内每本杂志的热点领域。并且对交叉领域做特别的表示。此图提供总体情况的overview。在该图的基础上，另外设计用户的交互行为，以显示更多的信息，包括details和zoom-in。
2. 设计一个用于统计的视图。可以显示每本杂志在当前一出现的所有研究领域上的研究比重。还可以各自的数据统计。
3. 之前所述的两个图之间的连接。设计页面提供两个图之间的切换，达到统一的展示效果。
4. 总体科学地图的设计

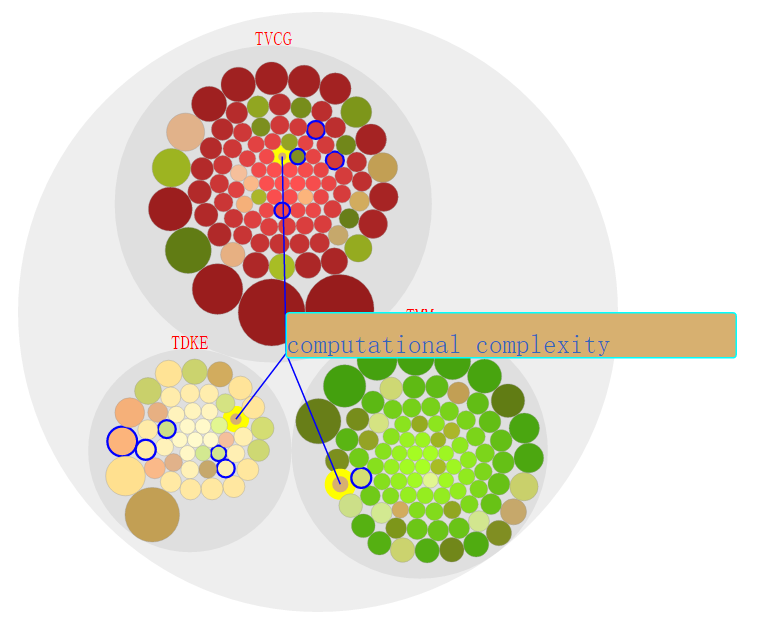


图3.1 2006-2011年总体科学地图

以及用户交互：领域名称及交叉领域的表示



图3.2 用户交互：某一领域的放大

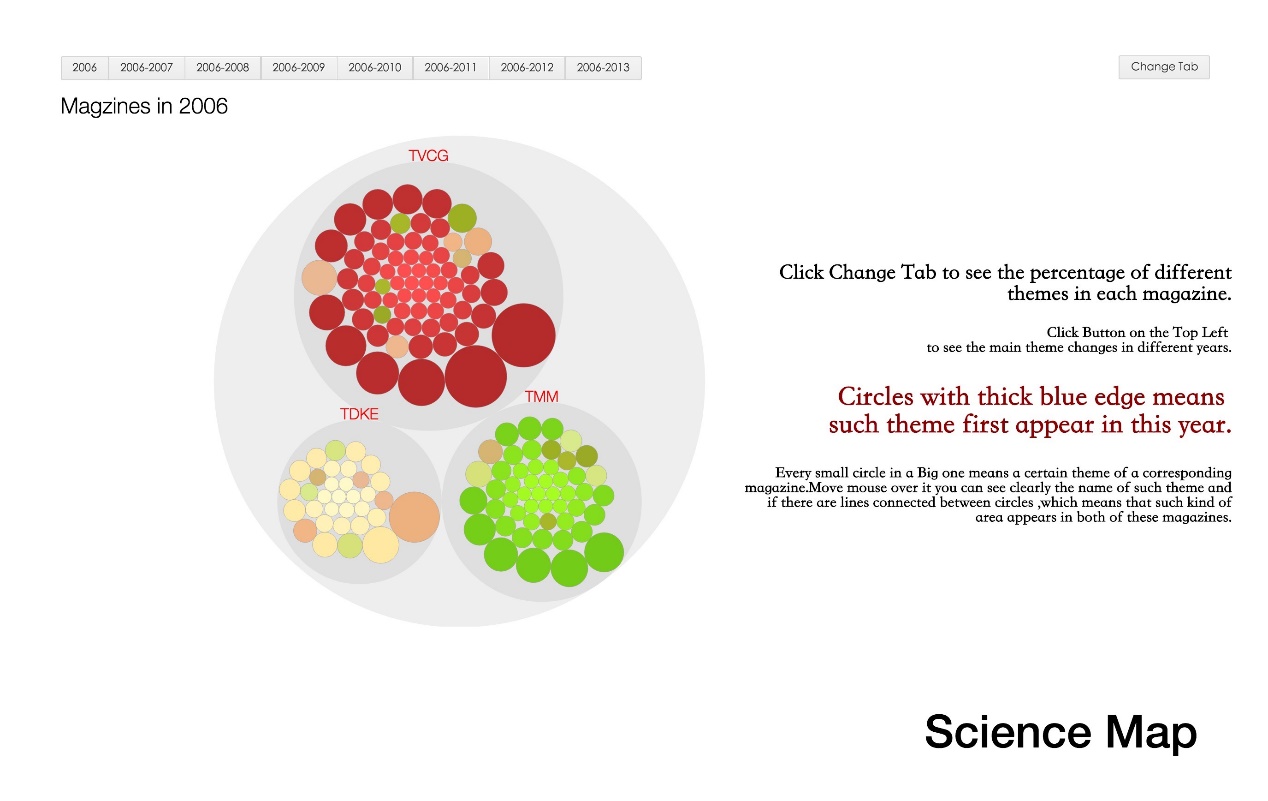


图3.3 用户交互：选择不同的截止年份2006

* 1. 静态科学地图Overview

我们采用pack图作为总体的科学地图。每本杂志对应一个大的圆，其中包含的小圆为该杂志中出现的热点领域。

某本杂志内不同热点领域之间存在着差别，即与该领域相关的论文数量的不同，也就是研究所占比重不同。对这样的差别，我们使用小圆的半径以及颜色深浅来表示。具体来说，如果某个杂志中对A热点领域的研究比重大于B热点领域。则在该杂志对应的大圆中，表示A的小圆半径要大于小圆B，并且颜色也要比B更深。

不同杂志的热点领域，除了所属大圆不同，还使用了颜色进行更加直观的区分。示例中，我们使用红、黄、绿三色来表示TVCG、TKDE、TMM独自出现在三个杂志中的热点领域。

而对于交叉领域的颜色表示，采用颜色叠加的方法进行处理。具体为，对于某一个交叉领域A，分别根据其在某本杂志的研究比重计算出颜色值，然后将多个颜色值进行加和（如果RGB的某个值超过255，则取255）。

另外，在某一个具体截止年限的地图中，应该对杂志中首次出现的热点领域进行标识。在图中，我们使用加粗的蓝色边框特别标识这些领域。

在图3.1中我们可以看到，在2006-2010年间出现在三本杂志中的研究领域。以TVCG为例，圆中深浅不同的红色表示的是仅出现在TVCG中的热点领域，而其他的颜色指示的是与另外两本杂志存在的交叉领域。圆中出现的四个蓝色边框的圆，表示出了在2010年，TVCG出现了四个新的热点领域。

* 1. 添加用户交互功能

交互1：在原图的视野中需要对某一个小圆表示的领域进行显示。用户可以使用鼠标指向某一个圆，来查看具体的名称。

交互2：对那些交叉领域，提供两者或三者之间的关联表示。在用户指向某一交叉领域时，对其他杂志中的相同领域进行高亮（黄色粗边框），并用蓝色的连线表示关联。如图3.1所示，三本杂志同时存在对computation complexity的研究。

交互3：zoom-in某一个热点领域，查看其详细的内容。对一个领域我们关注它的领域名称、首次出现年份，以及它在该杂志中的比重。如图3.2所示，点击TVCG中最大的圆，该圆被放大，可以看到首次出现的年份是2006年，其所占权重为118

交互4：截止年份不同的科学地图。用户可以通过页面上方的选择按钮，选取某一截止年份，来查看自起始年份到改年的科学地图。如图3.3所示，当用户选择2006-2011时，出现的新的地图。图中，新出现的领域（粗蓝色边界的小圆）以及每个领域所占比重（圆的半径）都得到了更新。

1. 用于统计的多维度圆图

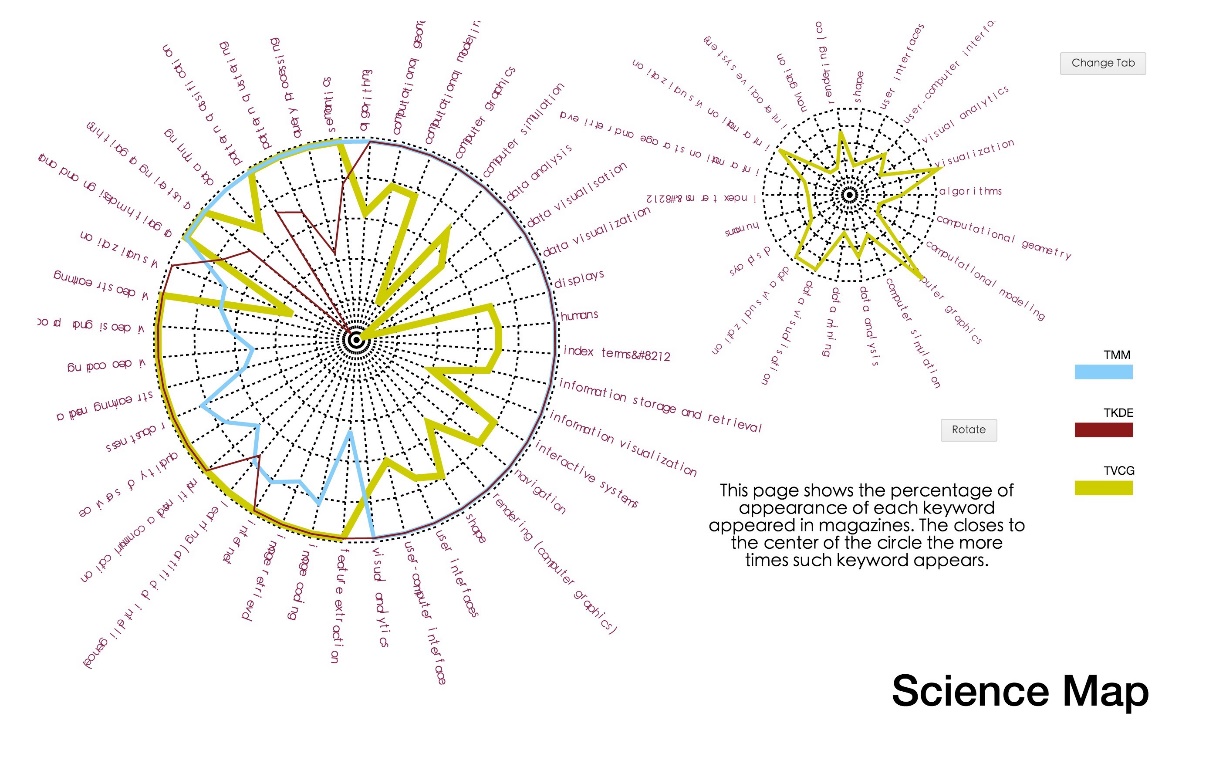


图3.4用于统计的多维圆图

* 1. 整体统计图

我们设计了一个多维圆图来显示各本杂志在多个领域的研究侧重比例。将三本杂志的专注领域作为一个集合，集合中每一个领域都是一个维度。杂志在每个维度上的取值为在该领域上的论文比例（侧重比例）。这样每一个杂志都对应一个多维的数据。

为了在一个图中显示多维的数据，我们采取了圆。将圆的角度均分，份数为选取领域的个数。在划分后的圆上的每个半径外侧标注一个领域。这样经过划分后的圆作为多维数据的坐标圆。之后将某一杂志在每个领域的侧重比例标在对应半径上（侧重比例沿半径从圆心向外递减）。将一本杂志在所有半径上的点连线，表示出该杂志的最终多维数据。不同杂志的曲线使用不同颜色来标注。

将三条曲线置于同一坐标圆中，得到整体的统计图。通过三条曲线在各维度下的交叉情况，可以分析他们之间的关联。

如图3.4所示，左侧的坐标圆中是整体的统计图。观察此图，可以发现，不同杂志在多数的维度下交叉现象不多，解释为每个杂志的专注领域各有不同。但也存在一些维度下的交叉现象，解释为各杂志间也存在交互的研究领域。

* 1. 添加用户交互功能

交互1：点击某一曲线，查看对应杂志的单独统计图。整体统计图上的每一条曲线可以被点击。点击后，出现该杂志的统计图。图中，一个维度（半径）上，数值递增的方向与整体统计图相反，即圆心处为0，向外递增。并且该图只显示该杂志的专注领域。也就是说，在单独统计图的所有维度中取值都不为0，即曲线不会深入到圆心的位置。该图可以帮助分析具体杂志的领域侧重比例，发现能够代表该杂志的领域。

如图3.4所示，右上侧的小圆是TVCG的单独统计图。可以发现，图中每个维度数值都不为0。并且在visualization和computer graphs上的比重最大。

交互2：允许图的旋转。由于每一维度的文字标注在圆的外侧，所以文字方向不一致，一些文字难以阅读。在页面中为用户添加rotate按钮，点击后圆可以进行顺时针的旋转，再次点击则停止旋转。这样用户可以方便的阅读每一个维度下的介绍了。

1. 上述两图间的关联

为了使上述两个视图之间的切换更方便。在页面的右上角添加Change Tab，方便两者之间的切换。

在图3.3总体科学地图中，点击Change Tab后会切换至图3.4统计视图，再次单击则回到总体科学地图

## 四、程序安装及部署：

将整个项目放入tomcat根目录下的webapp目录中，通过bin/startup.sh启动tomcat，在浏览器中输入localhost:8080/vis/web/pages/index.html访问