浙 江 大 學

# 本 科 生 毕 业 论 文
# 开题报告


学生姓名: 魏 昊

学生学号: 3073031107

指导教师: 周 昆 刘世霞

年级与专业: 07 级 计算机科学与技术

所在学院: 计算机学院

一、题目：社会化媒体信息可视化技术研究

二、指导教师对开题报告、外文翻译和文献综述的具体要求：

开题报告：对该课题的背景做深入的了解，明确课题任务、对课题进行可行性分析、制定研究方案以及课题进度计划。

文献综述：针对该课题调研相关文献，对社会化媒体、信息可视化技术做初步的了解；深入了解 Twitter 信息可视化研究，学习并研究标签云、时间轴、主题河等可视化基础技术。

外文翻译：翻译文献 A Visual Backchannel for Large-Scale Events。这是一篇综合性较强的文章，集中介绍了社会化媒体信息可视化概念、数据分析、多种信息可视化技术以及对可视化系统的分析。

指导教师（签名）_____

年　　月　　日

# 毕业论文开题报告、外文翻译和文献综述考核

**导师对开题报告、外文翻译和文献综述评语及成绩评定：**

该生论文研究的是社会化媒体信息可视化技术。开题报告整体组织逻辑感强，结构清晰，内容详实。通过文献综述和外文翻译较好地总结了前人的工作，并对课题有了深入的了解。从开题报告、文献综述和外文翻译的完成情况来看，该生深入了解了课题内容，并对可视化研究方法有了初步的探索。

| 成绩比例 | 开题报告<br>占（20%） | 外文翻译<br>占（10%） | 文献综述<br>占（10%） |
|---|---|---|---|
| 分 值 | 95 | 98 | 98 |

导师签名_____

年　　月　　日

**答辩小组对开题报告、外文翻译和文献综述评语及成绩评定：**

| 成绩比例 | 开题报告<br>占（20%） | 外文翻译<br>占（10%） | 文献综述<br>占（10%） |
|---|---|---|---|
| 分 值 | | | |

开题报告答辩小组负责人（签名）_____

年　　月　　日

-

# 目 录

# 本科毕业论文开题报告

## 1. 课题背景

近年来，**社会化媒体**（social media）在互联网的沃土上蓬勃发展，爆发出令人眩目的能量。社会化媒体现阶段主要包括博客、论坛、微博和社交网络等，它已成为人们彼此之间用来分享意见、见解、经验和观点的重要工具和平台。社会化媒体传播的信息已成为人们浏览互联网的重要内容，不仅制造了人们社交生活中争相讨论的一个又一个热门话题，更进而吸引传统媒体争相跟进。例如，现在 Twitter 的每日流量已经超过了 6 亿 5 千万条，人们能够在 Twitter 上跟踪相关人物的新状态、参与讨论热点话题。人们通过社会化媒体可以自由地交流对事件的感受、评论和建议；不仅如此，它还能够帮助参与者更深入了解事件乃至改变事件的结果。

自十八世纪后期数据图形学诞生以来，抽象信息的视觉表达手段一直被人们用来揭示数据及其他隐匿模式的奥秘。上世纪 90 年代期间新近问世的图形化界面，则使得人们能够与可视化的信息进行直接的交互，从而造就和带动了近十多年来的**信息可视化研究**。信息可视化试图通过利用人类的视觉能力，来理解和分析抽象信息的意义，从而加强人类的认知活动。信息可视化囊括了数据可视化、信息图形、知识可视化、科学可视化以及视觉设计方面的所有发展与进步。在这种层次上，如果加以充分适当的组织整理，任何事物都是一类信息：表格、图形、地图，甚至包括文本在内，无论其是静态的还是动态的，都将为我们提供某种方式或手段，从而让我们能够洞察其中的究竟，找出问题的答案，发现形形色色的关系，或许还能让我们理解在其他形式的情况下不易发觉的事情。信息可视化致力于创建那些以直观方式传达抽象信息的手段和方法。可视化的表达形式与交互技术则是利用人类眼睛通往心灵深处的广阔带宽优势，使得用户能够目睹、探索以至立即理解大量的信息。

目前展示社会化媒体内容的主要方式还是基于时间顺序的文本列表，而人们难以从海量的文本数据中提取出有效的信息，因此大大限制了接受信息、分析问题的能力。随着网络论坛、博客以及 Twitter 的广泛普及，文本可视化研究再度得到重视，新的问题及研究方向应运而生。信息可视化技术借助人类与生俱来的对图像信息的迅速辨识及分析能力，通过将原始数据转换成直观的、可交互的展现形式，使人们迅速获得有用信息，从而达到对数据进一步理解，分析及推理的目的。对于文本可视化技术的研究，自信息可视化领域诞生以来便从未中断过。近些年来随着社会化媒体的广泛普及，

及电子邮件等新兴沟通技术的广泛应用，文本可视化研究再度得到重视，新的问题及研究方向应运而生。针对文本可视化分析技术的研究主要包含三个方向：基于文本内容的可视化分析，基于文本关系的可视化分析，以及近两年来出现的，基于多层面信息(Multifaceted)的可视化分析。在毕业设计中，我主要实现了基于文本内容的可视化分析，它以文本内容作为分析对象，其主要任务包括内容总结、主题分析及可视化展现。通过基于文本内容的可视化分析，我们能够为用户将海量的数据进行系统性的分析和总结，并从中挖掘出密度高、有价值的信息，最终以美观、简洁的方式呈现给用户。

## 2. 目标和任务

**目标：**实现 Twitter 数据的信息可视化，依据原始数据进行主题分析，并实现可视化展现。可视化的展现方式能够使用户直观地了解一段时间内网络流行话题、热点话题随时间的进化过程、话题进化的细节以及推动话题进化的主要原因。

**任务：**本课题基于收集的大量 twitter 原始数据，首先进行数据的整理、数据结构归一、主题分析并建立索引，通过分析数据信息提取出关键话题，并实现三个方面的可视化表示。实现对不同时间区间的数据进行对比，并能够对大量数据进行处理。

1） 实现“主题河”（Theme River）的可视化表示：首先提取 Twitter 文本数据中的主题，然后按照时间对主题集进行分割，将主题看作在时间维度上延续不断的波流并展现为时间轴上弯曲的条带。不同的主题叠加在一起用边界或颜色区分从而生成一种统一的对文本流媒体的展现方式。“主题河” 在展现有效信息的同时，给人以美的享受，带来了更好的用户体验。“主题河”不仅能够使用户了解某一个主题的演化进程，还能够对数据集合整体轮廓有较为直观的认识。
2） 实现“标签云”（Tag Cloud）的可视化表示：抽取 twitter 数据的关键词并将其按照一定顺序，规律及约束整齐美观的排列在屏幕之上。这一可视化表示有助于人们了解近期热点话题。
3） 实现“时间轴”（Timeline View）的可视化表示：在时间轴上显示某一主题在一段时间区间的重要文本内容，有助于帮助人们分析该主题的进化、发展情况和近期关注热点。

# 3. 可行性分析

该信息可视化课题的可行性分析包括一下几个方面内容：

1）数据收集及访问可行性：与 MSRA 数据挖掘组进行合作，收集了 twitter 数据，首先进行了数据结构的分析和整理，由数据挖掘组的同事将 Tweet（用户发到 Twitter 上的信息）归一到统一的数据结构，并建立了 Tweet 的 Lucene 索引用于数据的高效访问。能够使用 Lucene 索引访问数据的 API。

2）技术可行性：本课题主要采用了 C#和 SilverLight 进行相关的开发，SilverLight 是面向对象的网页可视化开发工具，它的灵活性、安全性和易用性为可视化编程提供了良好的条件。基于数据挖掘组提供访问索引数据的 API 是 C#工程的动态链接库，还需要 WCF 通信技术。SilverLight 集成了 WCF 技术，能够通过 WCF 异步访问 C#类库，实现数据的读取操作。

3）应用可行性：本课题主要是借助信息可视化技术实现社会化媒体信息的直观、可交互的展现，使人们迅速获得有用信息，从而达到对数据进一步理解，分析及推理的目的。因此本项目可以为人们提供了话题演化进程的直观图像，以及热点话题的可视化展示，并进一步展现了话题的细节内容。课题还可以为社会话题的研究提供直观快捷的结果，将有广泛的应用。Twitter 数据为网络公开，未侵犯他人隐私，属于合法行为。

4）时间可行性：在冬学期便于导师进行沟通和讨论，基本确定了课题的主题，开始动工比较早，有足够的时间完成可视化设计，并最课题作进一步分析。

# 4. 研究方案和关键技术考虑

基于收集的 twitter 数据，首先建立 tweet 数据索引用于数据访问，在此基础上实现主题河、标签云以及时间轴的可视化表现。

## 研究方案：

1、题目名称：社会化媒体信息可视化技术研究

2、选题依据：当前社会化媒体内容的主要方式是基于时间顺序的文本列表，难以使用户从大量的文本数据中提取出有效信息，因此大大限制了接受信息、分析问题的能力。信息可视化技术通过将原始数据转换成直观的、可交互的展现形式，使人们迅速获得有用信息，从而达到对数据进一步理解，分析及推理的目的。可视化社会化媒体内容是近期可视化的热点，具有广泛的应用潜力。该课题将“主题河”、“标题云”和“时间轴” 三个可视化组件有机结合，为用户呈现统一、美观的可视化

展现。

3、研究范围：2010 年 Twitter 原始数据集

4、研究步骤、方法和进程

1）进行大量文献的阅读，积累社会化媒体和信息可视化等相关方面的背景知识。

2）与指导老师讨论后确定研究材料、可视化方式。

3）进一步阅读文献，查阅可视化方法前人的实现机制，深入学习可视化方面的内容。

4）数据的读取，数据 Lucene 索引机制及其 API，实现对原始数据 Lucene 索引的建立和访问。

5）实现“标签云”（TagCloud）布局，利用扫描线算法实现标签按照权重的摆放；

6）实现“时间轴”（Timeline View）展现，对某一话题选出代表性 Tweets，并按照时间顺序将其无重叠地展示在时间轴上。相应鼠标事件，提供 Tweet 信息细节。

7）实现“主题河”（Theme River）的可视化展现形式，按照时间对主题集进行分割，将主题看作在时间维度上延续不断的波流并展现为时间轴上弯曲的条带，展示主题随时间演化的过程。

8）撰写课题论文。

5、研究材料来源：微软亚洲研究院（MSRA）数据挖掘组收集的 Twitter 数据

6、研究的条件要求：Visual Studio 2010，SilverLight4，操作系统 Win7

7、研究的指导力量配备：MSRA 首席研究员刘世霞作为指导老师，指导本人完成该课题设计。

8、研究前成果预期及其表现形式：Twitter 信息可视化由网页形式展示。

## 关键技术考虑：

1）C#与 SilverLight 网页开发技术。

2）主题河：主题河的形状轮廓、层次排序、标签以及颜色选择；在主题河中，条分支代表一个主题。每个主题是由一系列沿时间轴的点集通过 Bezier 曲线的连接绘制而成。关于如何选取主题河的基线、如何排序不同的主题、标签的布局以及颜色的选取等方面都是决定该可视化优美、直观的重要因素，也是实现它的重要技术考虑。

3）标签云：标签云的布局（scatter plot）：实现标签位置无重叠，标签颜色、大小的确定。

4）时间轴：不同 Tweet 标签保证不重叠，并实现 Label 中文本自动换行（word wrapping）、由时间轴到 Label 连线的边线布局（edge routing）等技术。

5）WCF 通信：是由微软发展的一组数据通信的应用程序开发接口，它是.NET 框架的

一部分，集合了几乎由.NET Framework 所提供的通信方法。在课题中需要将C#动态链接库文件应用到 SilverLight 工程中需要 WCF 技术支持。

## 5. 预期研究结果

1） Twitter 信息可视化系统，包括主题河、标签云和时间轴等网页形式的可视化组件。

2） 课题研究论文。

## 6. 进度计划

- **（文献查阅）：** 2010 年 12 月至 2011 年 1 月
  查阅社会化媒体方面的文献，了解社会化媒体的发展历史、信息表达方式等相关信息，了解其数据结构和信息传递特点，如信息的转发、回复以及引用关键词组功能。查阅信息可视化方面的文献，了解信息可视化的发展历史及相关研究、应用。

- **（课题设计）：** 2011 年 1 月至 2011 年 2 月
  与导师讨论、完成对研究对象可视化方式的确定。

- **（数据处理）：** 2011 年 2 月至 2011 年 3 月
  将原始数据整合到同一数据结构并建立索引，熟悉通过索引访问数据的借口。

- **（研制开发）：** 2011 年 3 月至 2011 年 5 月
  实现数据的可视化，包括主题河、标签云和时间轴。

- **（撰写课题论文）：**2011 年 5 月至 2011 年 6 月

- **（结题和答辩）：**2011 年 6 月至 2011 年 6 月

---

* 注：目录供参考

# 本科毕业论文文献综述

**[摘要]** 社会化媒体（social media）近年来在世界范围内蓬勃发展，已成为人们交流的重要工具和平台。然而，目前社会化媒体展示内容的主要方式还是文本列表，这一形式难以使人们从大量的文本数据中提取有用的信息，大大限制了人们接受信息、分析问题的能力。信息可视化技术借助人类与生俱来的对图像信息的迅速辨识及分析能力，通过将原始抽象数据转换成直观的、可交互的展现形式，使人们迅速获得有用信息，从而达到对数据进一步理解，分析及推理的目的，对社会化媒体的进一步发展有着重要意义。

**[关键词]** 社会化媒体，信息可视化，微博，互联网

## [文献综述]

我毕业设计的课题是《社会化媒体信息可视化技术研究》。通过查阅并学习、分析相关文献资料，我逐步了解了信息可视化这一领域，并初步搭建了如何利用可视化方法来表示社会化媒体 Twitter 信息的框架。现对相关文献的有关内容做如下分析：

### 1. What is Twitter, a Social Network or a News Media?

本文主要目的在于研究 Twitter 的拓扑特征和它作为新兴信息分享媒介的能力，这是第一篇量化研究网络微博领域信息传播的文章。通过本文，我了解了 Twitter 上用户跟踪被跟踪、信息如何发布并被引用、用户如何相互交流信息的基本机制，并对 Twitter 的信息传递能力有了深入的认识。下面对文章主要内容进行概括：

文中首先介绍了 Twitter 数据采集的方法以及如何过滤垃圾信息。Twitter 官方提供了便于利用网络爬虫技术收集数据的 API，作者采用该 API 实现数据的采集，主要包括用户简历、跟踪与被跟踪的关系、热点话题及相关 Tweet。本章采用了 FireFox 浏览器插件 Clean Tweets 的机制过滤 Tweet 垃圾。

然后作者对收集的数据做了基础性分析，包括用户跟踪和被跟踪数量的比较、被跟踪人数与发布 Tweet 数量的关系分析、相互跟踪的百分比以及 Twitter 人际网络的分离度。Twitter 上单向跟踪的比例远远高于双向跟踪，由此可是说明比起社交网络媒体，Twitter 更像一个信息来源渠道。

文章用三种方法对用户做了排序：采用跟踪者的数量、PageRank 算法（用户关系网络）以及 Retweet 数量。跟踪者的数量和 PageRank 算法得到的结果相似，但是与按

Retweet 数量排序的结果相似度很小。这一分析说明 Twitter 中用户热门程度和 Tweet 内容热门程度并不是正相关，两者相对独立。

作者进一步对热点话题进行了分析。Twitter 与 Google 的流行话题相似性很小，交集大多为世界重大事件、名人以及电影；而且 Twitter 的流行话题维持的时间较长。Twitter 用户趋向于讨论新闻头条并回应新鲜消息。

文章最后分析了信息得以在 Twitter 上快速传播的途径——Retweet。通过统计分析，作者发现不论一个用户的跟踪者是多是少，一旦 tweet 开始通过 Retweet 机制传播开来，这条 tweet 信息的引用就能够达到一定数量。因此，Retweet 机制为每一个用户提供了将信息广泛传播的能力。

## 2. A Visual Backchannel for Large-Scale Events

本文主要介绍了一系列创新的可视化方法，用于跟踪并发掘社会化媒体中的重大事件。文中提到，当前社会化媒体仅以简单的文本列表形式来分享更新的内容，人们很难得到对事件演变进程总体、直观的认识。文中以 Twitter 数据为基础，提供了一系列进化的、交互的并且多元化的可视化方法，包括展现一个话题在一段时间内进化过程的“话题流”（Topic Stream），展示参与者以及他们的活跃性的“人物螺旋”（People Spiral），以及按知名度来展示与事件相关图片的 “图像云”（Image Cloud）。这三种工具为不断更新的文本列表信息提供可视化表达，并且他们自己也在不断的更新进化。这样方便了人们对信息的参与、查询和交互。文中还主要讨论了如何在数据动态更新的情况下实现进化式可视化表达的设计考虑。通过阅读与裂解本文，我初步学习了几种有效的可视化方法，将会参考“话题流”、“图像云”进行毕业设计系统的可视化表达。

作者首先阐述了 tweet 文本分析的过程。tweet 规定由不超多 140 个字母构成，并且除了 hashtag（Twitter 中用#开头，用于标记话题关键字的单词）之外不包含直接关键字。这对这一特点，作者采用了以下几个步骤来提取重要话题并降低噪音和不一致：

1）提取以超链接方式呈现的与事件相关的图片，并判断一个 tweet 是否为 Retweet。

2）对大量 tweets 进行聚类，移除特殊字符和超链接，并将文本转换成小写形式；然后将文本分裂为一个个独立的单词。

3）移除休止符和通用单词。

4）利用 Porter Stemmer 算法将意义相近的单词聚类（如名词、形容词、副词及单复数形式的聚类）。

5）将提取出的信息和原始 tweet 之间建立关系数据库。

然后文章主要介绍了设计 twitter 可视化的方法，包括话题流、人物螺旋和图像云。

1）话题流：

占据了可视化界面的大半部分，是一种可视化动态更新的原文数据并支持交互查询（平移和放缩）的堆栈图。选择它的目的在于它可以展现当前和一段时间内积累的话题演化情况。X 轴显示时间维度，Y 轴显示话题在 Twitter 上某一时间的相对频率。对于话题由上至下的排序，文中采用的是根据话题的发起时间决定：最近的话题放置在上面。色彩的选择上，新的话题为绿色、老话题为蓝色，并且热门的话题拥有相对较高的饱和度。话题流的形状采用立方 Bezier 曲线。每个话题还有一个标签用于注明该话题关键字。作者还实现了一系列展示话题随时间改变的动态方法，使其动态可交互。

2）人物螺旋：

由点和标签组成，用于展示话题的参与者以及他们之间的活动。螺旋中的每一个点代表一个参与者，点的大小代表他们的活跃程度，点的颜色代表参与者发布 tweet 的原创性。螺旋外围的标签标示参与者的用户名。这一可视化形式形象展现了话题参与者的动态，便于用户与热门话题参与者进行联系和交流。

3）图像云：

提供某一话题被分享的图片的预览。图像的尺寸由被引用的次数决定。图像的位置用一种迭代强制有向层算法决定，使尺寸较大的图像尽量靠近中部，尺寸小的图像靠近边缘。算法还对图像加了一定的旋转，使其达到美观的效果。这一可视化方式将事件的图像呈现给用户，利用人的视觉更为直观地展现事件内容。

## 3. Stacked Graphs – Geometry & Aesthetics

基于要实现 twitter 数据的“话题流”，我查阅了该文。这篇论文主要描述了“堆栈图”（“话题流”的原型）的算法和涉及决策，并且讨论了“堆栈图”在网络上获得的反响。文中认为这种复杂的层次化图形能够高效展示大型数据集的内容，并提供了具体绘制图形的数学分析。论文主要集中论述了“堆栈图”的四个设计决策因素：从优化总体轮廓的角度寻找绘图基线，排列图层的决策，如何放置文字标签以及为每一图层上色。通过阅读该文，我了解了实现“主题流”的大致方法，为毕业设计打下了算法基础。下面主要介绍下文中阐述的关于图像轮廓、排序、标签放置和着色的基本思路和方法：基线定义：话题的时间序列是由 n 个非负函数 $f_1, \ldots, f_n$ 组成的一个集合来表示，每个函数代表的一个话题。假设他们相异且定义在[0,1]区间上，堆栈图的基线 $g_o$ 被定义为图的底部。$g_i$ 为某一层图的顶部。基线定义有以下 4 种方式：

1）$g_0 = 0$，图层基于 y = 0 一层层向上叠加；

2）对称算法，使图层以 X 轴为对称轴对称，$g_o = -0.5(\text{sum}(f_i))$

3）减少图层总体轮廓的摆动（及全局斜率变化），$g_o = -1/(n+1)(\text{sum}((n-i+1)\, f_i))$

4）在 3 的基础上加了权重，进一步优化图层总体轮廓，利用数值积分进行计算。

-

着色：颜色的阴暗和饱和度代表一条话题流，话题的开始时间和受关注程度决定了其颜色。话题按照开始时间由冷色逐渐变为暖色，热点话题颜色深，非热门话题颜色浅。
标签放置：每个话题的标签放置在这一层最大空白处，尽量优化字体大小。在分辨率不够的情况下，使用鼠标动态显示标签的方法。
图层排序：采用根据权重由内到外排序，开始时间早的话题放置在中间，开始较晚的话题分布在外部两边。这样的放置方法能够使图层得以平衡，具有较好的视觉效果。

## 4. Interactive, Topic-based Visual Text Summarization and Analysis

本文主要介绍了一个交互式的文本可视化分析工具 TIARA，它集成了当前最先进的文本分析技术和交互可视化技术，用于可视化地总结大规模文本数据。该工具的主要特点一是利用 LDA 模型从文本集中抽取出话题，并输出一系列有关键字描述的话题以及它们的概率分布；二是提供了丰富的交互工具，如魔法透镜及与文本原文建立的链接。

通过阅读本文，我了解了一些文本分析总结的方法，包括基于语句的分析、基于关键字的分析等内容；还了解了信息可视化的方式，包括基于元数据（关于数据的数据）的可视化和基于文本内容的可视化。TIARA 系统形象展现了主题强度随时间变更的趋势、并用关键字描述了主题内容的细节，这一点与我的课题主干相似。作者从以下三个方面详细描述了基于话题的可视化文本总结方法：

1）基于话题的文本总结：TIARA 利用 LDA 模型在两个级别上总结本文集，①自动从整个文本集合中抽取潜在话题；②用关键字集合总结每一个文档。
2）基于话题的可视化文本总结：用堆栈图的每个色层来展现不同的话题内容，包括时间演化、话题关键字和主题强度。该内容还有两个延伸：①优化话题在堆栈中的顺序；②用关键字云填满每一图层。
3）基于话题的可视化文本分析：①根据需求提供话题细节，如用鱼眼（fisheye）技术放大话题的细节；②根据需求提东文本信息，如展示含有选中关键字的本文片段。

## 5. Understanding Text Corpora with Multiple Facets

这篇文章提出了一个数据模型，从四个基础概念方面进行文本的可视化：时间、分类、未结构化内容和结构化内容。为理解该数据模型，文章设计了一个将趋势图和标签云技术杂交的可视化系统，并为用户提供了多种交互化分析文本的方式。

我阅读本文的主旨在于了解标签云（TagCloud）的布局算法。算法利用迭代贪婪策略，首先计算出放置标签的边界轮廓以及轮廓中心，然后利用扫描线算法对标签进行

布局，尽量将标签放置在离中心较近、不与其他标签重叠的地方。当找到某一标签的放置位置后，更新轮廓及扫描线，然后处理下一个标签。扫描线算法高效实现了标签的云布局，是我值得参考的方法。

## 6. 小结

上述几篇文献对我有了很大的启发，使我了解了 Twitter 社交网络的机制和数据分析的方法、可视化领域的方法和算法基础。总而言之，查阅文献资料使我对课题有了系统性的认识，对我撰写毕业论文起到了很大的帮助作用，让我受益匪浅。

## 7. 参考文献

[1] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a Social Network or a News Media? Proceedings of the 19th International World Wide Web (WWW) Conference, April 26-30, 2010, Raleigh NC (USA)

[2] Marian D ¨ ork, Daniel Gruen, Carey Williamson, and Sheelagh Carpendale. A Visual Backchannel for Large-Scale Events. November/December 2010 (Vol. 16, No. 6) pp. 1129-1138 1077-2626/10/$26.00 © 2010 IEEE

[3] Lee Byron & Martin Wattenberg. Stacked Graphs – Geometry & Aesthetics. NOVEMBER/DECEMBER 2008 (Vol. 14, No. 6) pp. 1245-1252 1077-2626/08/$26.00 © 2008 IEEE

[4] Liu, S. , Zhou, M. , Pan, S. , Qian, W. , Cai, W. , & Lian, X. . (2009). Interactive, topic-based visual text summarization and analysis. Conference on Information and Knowledge Management, 543-552. ACM Press. Retrieved from http://portal.acm.org/citation.cfm?id=1645953. 1646023

[5] Lei Shi, Furu Wei, Shixia Liu, Li Tan, Xiaoxiao Lian, Zhou, M.X. Interactive, Topic-based Visual Text Summarization and Analysis. Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on

# 本科毕业论文外文翻译

**[文献来源]：** Marian D ¨ork, Daniel Gruen, Carey Williamson, and Sheelagh Carpendale. **A Visual Backchannel for Large-Scale Events**. November/December 2010 (Vol. 16, No. 6) pp. 1129-1138 1077-2626/10/$26.00 © 2010 IEEE

**[正文]**

# 大型事件的可视化反向渠道

## [摘要]

本文主要介绍了“可视化反向渠道”这一概念，它是一种新兴的、用于跟踪并探索关于大型事件在线会话的方式。微博社区（如 Twitter），作为一种人们交换对政治演讲、体育竞技、自然灾害等重大事件的简短评论和观点的数字渠道，正在以惊人的速度日益壮大。目前，微博中信息的更新分享还是以简单的列表形式呈现，这使得在高速讨论事件的同时难以为用户呈现事件的全貌，而且展现事件的演化过程也变得相当困难。为了解决这一问题，我们的可视化反向渠道设计提供了一种具有进化性、交互性和多方位的信息概述方式，用于呈现 Twitter 中大规模、发展中的会话。为了可视化连续更新的信息流，文章讨论了对正在发生的和已发生的事件进行可视化的方法。作为充分基于网络坐标视图的系统，文章主要介绍了以下几种可视化方式：①话题流（Topic Stream），一种基于时间且可调节的、利用堆栈图进行可视化话题的方法；②人物螺旋（People Spiral），用于展现会话参与者以及他们的活跃性；③图像云（Image Cloud），展示与事件相关的图像，并用图像大小表示图像的受关注程度。再加上会话本文列表，这些相互链接的视图支持话题、参与者和时间区域的交叉过滤。文中进一步讨论了系统的设计考虑，特别是如何以进化的方式可视化展现动态更新数据。

**[关键词]：** 反向渠道，信息可视化，事件，多维视图，微博，信息检索，万维网

## [引言]：

数字反向渠道是一种正在逐渐成形的社会现象。人们在观看政治争论、参加教育活动或者面对自然灾害的同时，通过数字反向渠道分享简洁而实时的信息的活动频率

正在日益增长。这一交流形式创造了关于社会重大事件持续不断的会话。数字反向渠道已经成为一个引人入胜的交流媒介，越来越多的人在此平台上交流对事件的观点、建议和评论。数字反向渠道不仅能够让参与者分享经验和形成事件的观点，而且能够帮助参与者去影响事件的演变和结果。

当数字反向渠道在社会化信息空间中逐步提升自己重要性的同时，我们也看到了当前它所使用展示信息的方法的局限性。仅仅使用按照时间顺序的文本列表不足以充分展示大型反向渠道的信息，因为它们不能够形象地呈现实时会话的规模和运动。参与者因此无意地而又不可避免地对事件主题的认识产生了一定的偏差，并缺乏对反向渠道所涵盖信息的的总体认识。

为了探索这些问题，我们在文中介绍了可视化反向渠道，它是一个具有进化性、交互性和多方位的可视化接口，它集成了三种可视化工具并包含了通过链接列出的数字反向渠道。为了给数字反向渠道会话提供一个新的视图，文中我们介绍了话题流（Topic Stream），一种实时的、与话题相关的、可调整的堆栈图，它可以用于可视化从数字反向渠道的会话中提取出来的话题。于此同时我们还提供了两种紧凑的可视化工具，人物螺旋和图像云。它们集中展现了活跃的参与者和通过会话被分享的图片。这些可视化工具在话题发展过程中为当前的会话提供了直观的视觉感受，通过视觉加重使会话的展现具有丰富的视觉体验。通过这一方式，可视化反向渠道为正在发生的事件和时间上下文内事件演化的情况提供了可视化依据。这四个通过高亮、加粗和过滤连接起来的视图是为了能够达到以下要求而设计的：为连续更新的数据集提供有机进化的表达，交互式地访问按照时间分布的话题，查找话题中活跃的参与者，以及展示重要的图像。

该论文的贡献主要有以下两点：

1）提出了进化性可视化这一概念，它用连续的信息流集成地展现了当前活动和近期发展的情况，如数字反向渠道。
2）引入了三种新颖的交互式可视化工具，用于总结大型反向渠道的主要方面并提供在可视化渠道上下文中基于时间、话题和人物的探索互动。

论文的组织形式如下：第二章对相关工作做了一个简要的概述。第三章简述了探索更好数字渠道的动机，并对系统提出了设计目标。第四章阐述了处理动态、进化性数据的挑战和机遇。第五章详细描述了系统的实现，第六章讨论了系统最初得到的用户反馈。第七章提出了工作的局限行并指出了未来研究工作的大致方向，最后第八章对本文做了最后的总结。

## [相关工作]

我们的工作可以用于可视化持续会话、在广泛的事件中体验数字反向渠道，并表达话题在一定时间范围内的演化过程。

**可视化持续会话：**

以前多数关于可视化人类通信的研究侧重于考虑社会和表达的方面。例如，从较低层次上看，对输入样式、正确性和停顿的艺术性可视化可以丰富普通的基于文本的信息。从较高层次上看，可视化聊天参与者的用户名和在线状态、以及他们当前的活跃性和结构等，可以提高社会的关注度。可视化短期和长期的讨论能够提供关于会话活动和团体组织模式有价值的线索。进一步的研究致力于会话线程上。例如，论坛讨论的范围及某一术语的查询分布能够通过可视化论坛线程而得到整体的概貌。邮件交换可被可视化为代表回复结构和时序的线程弧线。虽然嵌套和时序为结构化、顺序化信息交换的展示提供了合适的机制，我们还是致力于为当前讨论的主要话题提供概览。

与会话的社会化、结构化等方面相反，几乎没有可视化工具试图去表达当前讨论的实际话题。会话地图（Conversation Map）通过坐标系视图接口展现了参与者、重要的话题以及新的讨论线程，它包括社会化的和语义的网络可视化。通过将关键字列表分布在水平时间轴上，Themail 对私人邮件会话中话题的发展过程进行了可视化。会话地图的 Themail 都提供了基于文本内容的分析，但是它们仅能够用于分析存档记录，不能分析实时会话。除数字会话之外，可视化接口可以展现参与者的活跃时间或用术语总结话题聚类。虽然前者总结了话题参与者及参与时间，话题的实际内容却无法展现出来。话题聚类的可视化缺乏先前技术所包含的历史记录可视化。

我们的工作延续了关于强调社会性和结构性方面讨论的研究，并探寻如何按照时间发展的趋势表达会话主题的同时展现参与者的活跃度和图像信息。前人的工作仅能狗展示小型在线会话或大型存档文本，而我们的工作实现了大型实时会话的可视化表达。

**将数字反向渠道用于事件：**

数字反向渠道可被理解为伴随着正在发生事件的会话。早期的研究将反向渠道通信定义为私人的、非正式的、短暂的且仅对参与者可见的交流渠道。接下来的工作开创了了公众在学术会议期间通过网络聊天这一方式，并发明了为观众的提问投票的交

互式系统。数字反向渠道的优势包括可以再不中断事件发展的同时交流附加信息、以及形成事件的能力。主要的担忧在于对话题潜在的扰乱和分散。

由于微博的出现（特别是 Twitter），我们正在目睹一种伴随重大事件以及私人事件的新形式反向渠道。研究在 Twitter 上发布的消息（tweets）表明，个人信息的更新、信息链接以及实时的事件评论正在被人们广泛地使用。研究者们已经注意到了 Twitter 的易访问、简洁、可移动、似广播、实时行以及用松散链接分享 tweets 等独一无二的特性。反向渠道同时还可以帮助学习者获得知识，这一点已被那些介绍反向渠道在教育和商业中的作用的书籍证实。

数字反向渠道作为信息交流渠道，正在被越来越多的人接受。通过对危机和高安全性事件的分析研究表明，与私人信息传播相对比 Twitter 更利于信息的散步。反向渠道通过自底向上地提供信息这一机制，被认为是与官方信息互补的消息发布源。

政治领域是数字反向渠道的另一个不断增长的热点。例如，用户在 2008 年美国总统竞选中通过 Twitter 展开了对候选人的激烈讨论和评估。通过 Twitter 引起的反应可以互补其他媒介可视化地展现出话题的趋势。由于数以万计的人参加了数字反向渠道的会话，跟踪话题讨论得发展一时间变得很困难。我们的工作致力于这一问题，尝试可视化表达绝大多数人的意见以及话题随时间的进化趋势。

**表达话题的改变：**

接下来我们针对信息分享的实时性和社会性讨论可视化研究。最具有相关性的早期工作是 ThemeRiver，它能够将用户选取的主题可视化为以水平轴为中心的堆栈图。近期通过将 ThemeRiver 中的方法应用到娱乐界的数据集中，人们找了新方法对流图进行排序和着色。堆栈图技术的主要约束在于限定了时间范围、主题的静态选取以及缺乏放缩和过滤操作。我们将会在工作中通过可视化在线更新数据以及允许依据话题、人物和查询项的放缩和交互式过滤，呈现一种实时的、动态的堆栈图技术。

可视化网络上人们交流中主观的、社会化的方面正在成为一大热点。许多基于网络的可视化样例可被称为信息可视化框架，它们是拥有广泛用户人群以及不同洞察类型人群的可视化实践。如一些可视化界面提供了用户和话题的网络可视化，另一些将关键字映射到地形图上。

总而言之，已有许多有为的研究是基于可视化话题的发展历程。在基于仅使用静态数据集、时间范围和话题粒度固定的堆栈图基础上，我们引入了支持高动态、高交互的堆栈图技术。我们希望在动作结束后加入新方法来表达近期数据变更，从而将工

作得以扩展。这是第一篇致力于将话题当前内容和过去改变内容相结合的进化性视图。我们希望通过可视化和交互式表达集成当前、近期以及过去话题改变的概述。

## [问题空间和范围]

在研究工程的开始我们同 20 多名同事一起进行了自由讨论，收集了他们期望数字反向渠道的初始反馈以及他们遇到的问题。在他们的反馈信息基础上，我们确定了一个范围较广的时间并选取了反向渠道已存在的主要问题。问题集中在如何跟踪正在被讨论的话题、了解讨论参与者以及对主要事件和反向渠道的管理注意。

信息过于丰富：由于活跃的事件和反向渠道拥有许多的参与者，这使跟踪会话或进行有意义的评估变得非常困难，因为有太多的发言和发言者。因此听从大多数人的意见是一个好的决策。

陌生人的海洋：在大规模时间中，由于很难得知其他的参与者，用户会感觉迷失在陌生人的海洋中。比如，从反向渠道中找到和自己想法类似的人是一件很困难的事。体验大型事件为认识新朋友提供了机会，而这一点要求事件的参与者能够对其他参与者可见。

失去焦点：分散性也是反向渠道的一个关键性难题。由于有太多的人物和信息可以跟踪，参与者很难决定选择什么跟踪，并同时缺乏对事件本身的关注。

## [设计目标]

为了能够达到反向渠道和主要事件之间的切换，我们试图在“最近”的上下文中表现“当前”。通过对近期发展的表达展现当前的活跃性，我们可以帮助参与者再造意识。

特别地，我们想帮助参与者回答正在进行的会话的普遍性问题，如现在主要的话题是什么？话题如何随时间而演化？最活跃的参与者是谁？某几个参与者正在谈论什么？基于这些问题、前人的研究以及我们自己的实验，为可视化反向渠道提出了一下设计目标：

1）对会话进行总结：可视化反向渠道应该包括总结反向渠道会话实时、社会以及图像方面的视图，减少跟踪反向渠道需要的认知过程。会话中主要话题、最活跃的

参与者以及受欢迎的图像应该被可视化，以便表达会话发展的进程。

2）集成“当前”以及“最近”：可视化应该通过当前和过去一段时间的发展来捕获反向渠道的演化过程。参与者能够通过对事件最近发展的了解而对当前的活动有更为直观的了解。

3）对当前进行扩展：因为“现在”非常短暂，可视化的重点应该放在当前的活跃性上，如正要来临的发言和可视化中随后的改变。这种方式可使当前的反向渠道活动更加利于理解。

4）提供灵活时间窗口：考虑到话题可能会持续几分钟到几周，界面应该提供灵活的方法用于修改时间窗口。时间区间的选择将会影响可视化改变的范围。时间间隔越窄，更多的活动被期望可视化。

5）使可视化活动可调整：当展现当前事件时，界面应该允许用户通过确定时间窗口、根据话题和参与者过滤以及键入搜索项来调整可视化活动。我们期望更少的可视化活动、更大的时间窗口以及更多的过滤约束。

6）创建有机的感觉：我们期望可视化反向渠道能够提供动态且吸引人的界面。色彩、形状和布局应该要达到易理解、吸引人并给人带来视觉愉悦等要求。由界面改变引起的数据更新和交换过滤需要通过动画过滤实现。

## [Tweets 数据]

我们用微博 Twitter 的数据作为数据源。除了 Twitter 被广泛使用这一原因外，Twitter 还提供了公开的 API 可以简化对公众 Tweet 信息的访问。

短实时更新：Tweets 是用户在 Twitter 上发布的信息，它已经成为信息可视化研究领域关注的热点，因为它具有简洁性、社会性、实时性和公众性。在使用初期为了与 SMS 文本兼容，Tweets 的长度被规定在 140 字符内，这一限制现在也可以被看为管理注意力的方法。我们致力于关注 Tweet 的作者、时间标签和文本内容。

Tweet 的文本有丰富的附加信息。文本可以集中表达会话内容，并在与时间标签结合的情况下表示话题的进化情况。Tweet 还可以包含图像链接（图像可来源于手机的拍摄），因此能够以图像形式展现事件的参与者。不仅如此，Tweet 用 RT 符号来表示 tweet 内容是引用自其他用户。

另外 Twitter 还提供了 hashtag 机制，如#visweek，用于在 tweet 中引用地点、话题、事件等专有名词。Hashtag 已成为一种重要的组织机制，它允许 Twitter 用户根据特定的 hashtag 跟踪 tweets 子集，及由围绕某一特定的 hashtag 构成的 tweet 子集。据统计，

17%的 tweet 包含 hashtag，每条 tweet 所包含 hashtag 的平均数为 1.3 个。我们可以使用 hashtag 或搜索关键字来继承属于某一事件的 tweet。在文本分析中，我们将 hashtag 看成单词。

## [文本分析]

我们所分析的文档有 140 字符长度的限制，并且不包含显式关键字、标题或除了 hashtag 之外的标签。我们用参与者在 tweet 中多次重复的单词来代表话题，相信可以有力表现可视化反向渠道的会话。用自由文本取代预选取的分类，会使话题层面的结果有较高的噪音和变种。以下几个处理步骤的目标在于减少噪音并从 tweet 中抽取有意义的话题：

1）图像和 Retweet：首先我们抽取出 tweet 中链接到图像服务器（如 Twitpic）的事件照片，然后查找 RT 符号，判断信息是否引用自其它人。这一操作提前于以下几个步骤，因为链接和 RT 不考虑在话题流中。
2）字符串清理：由于我们集成了大量数据，噪音是不可避免的。我们先移除特殊字符和超链接、并将文本转换成小写，得到的字符串仅由字符和数字组成。然后用空格作为分隔符将字符串分割，以便得到独立的单词。
3）停顿词移除：为减少无内容的单词，我们移除了 120 个停顿词和通用词（如”a”，”is”，“can”等）。我们也将作为搜索项的单词除去，因为它将与所有时间的 tweet 相关联而又没有增加任何信息。
4）词干：我们进一步用波特词干算法（Porter stemmer algorithm）合并带有相同意义的单词，从而减少单词量。比如，单词的单复数形式和名词、形容词、动词形式都将被合并为一个词。
5）联合：我们将抽取的信息和原始 tweet 本身存入关系数据库。总体上讲现在对每一个 tweet 有以下信息：发布时间、原始文本内容、抽取的信息、作者名称、图像链接以及它是否为引用等。

## [可视化反向渠道设计]

这一节我们主要讨论以坐标系视图接口设计可视化反向渠道，为当前活跃的、正在发展的反向渠道会话提供交互的可视化方法。除了传统的反向渠道信息列表，该可视化反向渠道界面由三个新颖的可视化工具组成，并在它们之间建立相互的链接以便

支持交叉过滤和高亮。

话题流：是可视化反向渠道的主要视图，它是一个用于可视化文本数据在某段时间内的动态改变的可交互的堆栈图，并通过支持时间的放缩、平移和话题过滤支持交互式查询。以前的堆栈图技术主要依赖于固定的时间间隔和预定义分类，如新话题、电影名称及艺术家姓名等。与它们相比，话题流支持多级时间放缩而且是基于不断更新的 tweet 文本内容。

人物螺旋：为了达到美观地展示参与者这一目标，我们设计了由点和标签组成的人物螺旋布局，从而显示了反向渠道参与者和他们的活跃性。螺旋中的每个点代表一个参与者，点的大小代表他们在反向渠道中针对某一话题的活跃度（由当前时间窗口内的相对 tweet 数量决定）。点的颜色表现用户发帖的原创性。如果用户引用得越多，点的饱和度越低。颜色的基色为蓝色。这一表达可以展示参与者在反向渠道中对新信息的贡献度。

图像云：图像云是基于网络上广泛使用的标签云（tag cloud）思想创建的，它是社会共享图片的轻量级视图。标签云用字体大小表现标签的重要性，我们通过设置图像的大小来指示该图像相对于其它图像被分享的频率。因为 twitter 信息中可以包含图像链接，图像和照片在事件上下文中有重要的社会意义。

## [总结]

本文中我们展示了可视化反向渠道，一种新颖的探索事件大型会话的媒介。它使以下方面成为可能：

1）得到一段时间内反向渠道会话的可视化表示；
2）跟踪实时、连续改变的数据集的进化表达；
3）探索话题的实时性、话题性、社会性和图像等方面。

为了可视化地总结反向渠道会话演变的过程，我们介绍了话题流、人物螺旋和堆栈图。这三个可视化表达为不断更新的信息列表提供了上下文环境，并且它们自身也在不断地更新和进化。用户能够通过选取时间间隔、参与者和兴趣话题将注意力集中到反向渠道会话的子集上。