第9章 文本和文档可视化

1 文本可视化释义

文本可视化释义

- 背景
  - 需要更高效的文本阅读和分析方法
- 核心问题之一
  - 辅助用户准确地从文本中提取信息、简洁直观展示信息
- 文本可视化应用广泛
  - 标签云技术，网站展示关键词的常用技术
  - 信息文本图，美国纽约时报等各大纸媒辅助用户理解新闻内容的必备方法
- 文本可视化与其他领域相结合
  - 信息检索技术，通过可视化来描述信息检索过程，传达信息检索的结果

文本信息的层级

- 文本信息需求多样，要求从不同层级提取与呈现文本信息
- 词汇级(Lexical Level)信息
  - 语义单元信息。语义单元(Token)，由一个或多个字符组成的词元，是文本信息的最小单元
  - 词汇级可提取的信息
    - 文本涉及的字、词、短语，以及它们在文章内的分布统计、词根词位等相关信息。常见，文本关键字
- 语法级(Syntactic Level)信息
  - 基于文本的语言结构，对词汇级的语义单元进一步分析和解释而提取的信息
  - 语义单元的语法属性，属于语法级信息
    - 如词性、单复数、词与词之间的相似性，及地点、时间、日期、人名等实体信息。通过语法分析器识别
- 语义级(Semantic Level)信息
  - 语义内容信息和语义关系，是文本的最高层信息
    - 分析词汇级和语法级所提取的知识在文本中的含义。如文本的字词、短语等在文本中的含义和彼关系
    - 及，作者通过文本所传达的信息，如文档的主题等

文本可视化的研究内容与任务

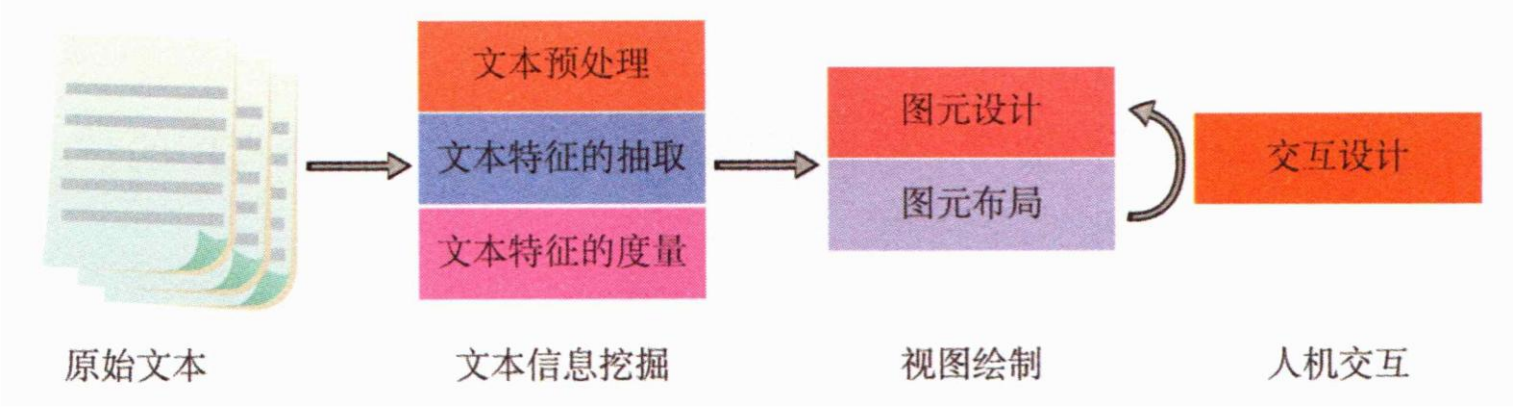
- 研究内容
  - 文本文档的类别
    - 单文本、文本集合和时序性可视化
  - 文本文档的内容
    - 文学作品、在线社交媒体、科学论文/专利、在线沟通数据、电影影评、医疗报告可视化等

文本可视化的研究内容与任务

- 分析任务（不同研究内容需要不同的分析任务）【描述高层次的用户分析需求】
  - 对文本的内容、特征进行总结(包括对词汇、句法的分析)
  - 分析文本中讨论的主题
  - 研究文本中蕴含的情感
  - 挖掘文本中描述的事件
  - 关联分析多源文档数据的内容、特征
- 可视化任务【描述相对低层次的展示和交互需求】
  - 找到文档中用户感兴趣的内容
  - 对文本的内容、特征进行聚类或分类
  - 比较文档和文档集合的各种信息
  - 查看文本内容的总体概览
  - 对文本内容，进行浏览与探索
  - 对文本中的不确定性进行分析

文本可视化流程

- 文本可视化的工作流程
  - 文本信息挖掘，视图绘制，人机交互



文本信息挖掘

- 文本数据的预处理
  - 过滤文本中的冗余和无用信息，提取重要的文本素材
- 文本特征的抽取
  - 采用文本挖掘技术，提取任务所需要的特征信息
    - 如，词汇级的关键词、词频分布，语法级的实体信息，语义级的主题等
- 文本特征的度量
  - 用户可能对（在多种环境下或从多个数据源）所抽取的文本特征的深层分析感兴趣
    - 如，文本主题的相似性、文本分类等

视图绘制

- 视图绘制
  - 将文本挖掘所提炼的信息变换为直观的可视视图
    - 在直观的可视图元的辅助下，用户可以快速地获取信息
  - 涉及图元设计，图元布局方法

人机交互

- 人机交互
  - 用户如何生成视图和满足分析需求而操作视图

2 文本信息分析基础

分词技术和词干提取

- 分词(Tokenization)
  - 将一段文字划分为多个词项，剔除停词，从文字中提取出有意义的词项
- 词干提取(Stemming)，去除词缀得到词根，得到单词最一般写法
  - 避免同一个词的不同表现形式对文本分析带来干扰

- “I have a dream that one day this nation will rise up and live out the true meaning of its creed: ‘We hold these truths to be self-evident,that all men are created equal.”
  - 经过分词后，可提取出20个词项：I, have, dream, one, day, nation, rise, up, live, out, true, meaning, creed, hold, truths, self-evident, all, men, created, equal
  - a,the,that等对文本语义影响较弱的停词已经被剔除
  - 接下来，词干提取过程将“men”和“truths”分别还原为“man”和“truth”

词汇级模型

- 向量空间模型(Vector Space Model)：利用向量符号对文本进行度量的代数模型
  - 指代一系列向量空间的定义、生成、度量和应用的方法与技术
  - 常用于自然语言处理、信息检索等领域

- 词袋模型
- 文本相似性度量
- TF-IDF

词袋模型(Bag-of-words Model)

- 词袋模型，用来提取词汇级文本信息。在过滤掉停词等对文本内容影响较弱的词之后，词袋模型将一个文档的内容总结为在由关键词组成的集合上的加权分布向量
- 在基于词袋模型计算的一维词频向量中，每个维度代表一个单词；每个维度的值等于单词在文本中出现的统计信息，可引申为重要性；单词间没有顺序关系
- “I have a dream that one day this nation will rise up and live out the true meaning of its creed: ‘We hold these truths to be self-evident,that all men are created equal.’
- I have a dream that one day on the red hills of Georgia,the sons of former slaves and the sons of former slave owners will be able to sit down together at the table of brotherhood.
- I have a dream that one day even the state of Mississippi,a state sweltering with the heat of injustice,sweltering with the heat of oppression,will be transformed into an oasis of freedom and justice.I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character.”

词	I	dream	color	skin	nation	slave	injustice	owner
词频	4	4	1	1	2	2	1	1

文本相似性度量

- 夹角余弦值（等），度量文本语义的相似度

$$\cos\theta = \frac{v_1 v_2^T}{\|v_1\| \cdot \|v_2\|}$$

  - 其中v_1 和 v_2 是两个文档的特征向量， $\|v_2\|$ 和 $\|v_2\|$ 是向量的模
  - 函弦值越大，表示这两个文档的内容越相似，反之亦然

TF-IDF(Term Frequency-Inverse Document Frequency)

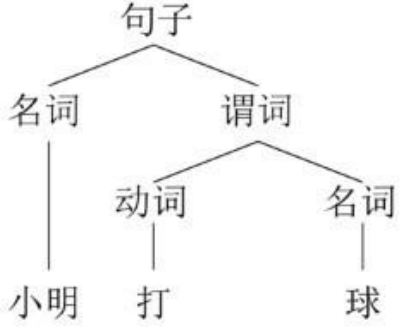
- TF-IDF，文档中的每个词合理地分配权重最常用的方法
  - $Tf(w)$：词 w在文档中出现的次数,
  - $Df(w)$：是文档集中包含词 w 的文档数目， N 代表文档的总数。
  - 定义

$$Tf\_Idf(w) = Tf(w) * \log\left(\frac{N}{Df(w)}\right)$$

- 代表词w对于某个文档的相对重要性
  - 和对词的重要性的直观认识一致，如果一个词对于某个文档越重要，那么
    - 它就越多地出现在该文档中($Tf(w)$值较大)
    - 并且越少地出现在其余的文档中($Tf(w)$值较小)

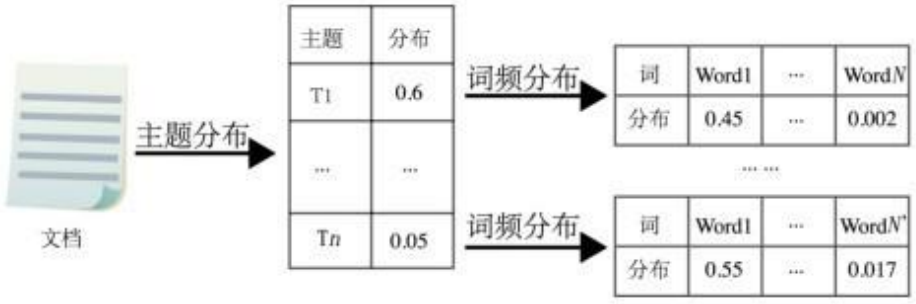
树结构模型

- 语法分析树(分析树，Parse Tree)，一种用于反映文本语句及其语法关系的有序树结构
  - 可按照短语结构语法(Phrase Structure Grammar)或依存语法(Dependency Grammar)中的依赖关系来生成
- 常用的语法分析树构建算法，分为自顶向下和自底向上
  - 自顶向下算法，从树结构的根节点开始构建，包括上下文无关语法、概率上下文无关语法、Earley算法等；
  - 自底向上算法，从树结构的叶节点开始构建，包括Cocke-Kasami-Younger算法等



语义模型

- 主题模型，从语义级别描述文本集合内各个文本的语义内容，即文本的主题描述
  - 文档的语义内容，可描述为多个主题的组合表达，主题可认为一系列词的概率分布或权重分布



- 文本主题的抽取算法，大致可分为
  - 基于矩阵分解的非概率模型
    - 词项-文档矩阵被投影到K维空间中，其中，每个维度代表一个主题。在主题空间中，每个文档用K个主题的线性组合表达而成。
  - 基于贝叶斯的概率模型
    - 主题被看成多个词项的概率分布，文档理解为多个主题的组合而产生。一个文档的内容是在主题的概率性分布基础上，从主题的词项分布中抽取词条而构成。

词嵌入

- 词嵌入
  - 自然语言处理中【语言模型和表征学习技术】的统称
  - 将每个单词表示为一个高维向量(常见的表示包括独热表示、分布式表示等)，再将这个高维向量嵌入一个低维连续向量空间中
- Word2vec
  - 一种用来实现词嵌入的工具，它依赖Skip-grams或连续词袋(Continous Bag of Words,CBOW)模型来建立神经网络

3 文本内容可视化

基于关键词的文本内容可视化

- 关键词可视化，以关键词为单位可视地表达文本内容。常见词频

- 标签云(Tag Cloud，Text Cloud、Word Cloud)，抽取文本中的关键词，并将其按照一定顺序、规律和约束整齐美观地排列在屏幕上
- 标签云利用颜色和字体大小反映关键词在文本中分布的差异

Wordle

- 用户可自定义画布填充区，正方形、圆形或花瓶形状等
- 既满足画布的约束，又提高空间利用率，Wordle改进了关键词的布局算法

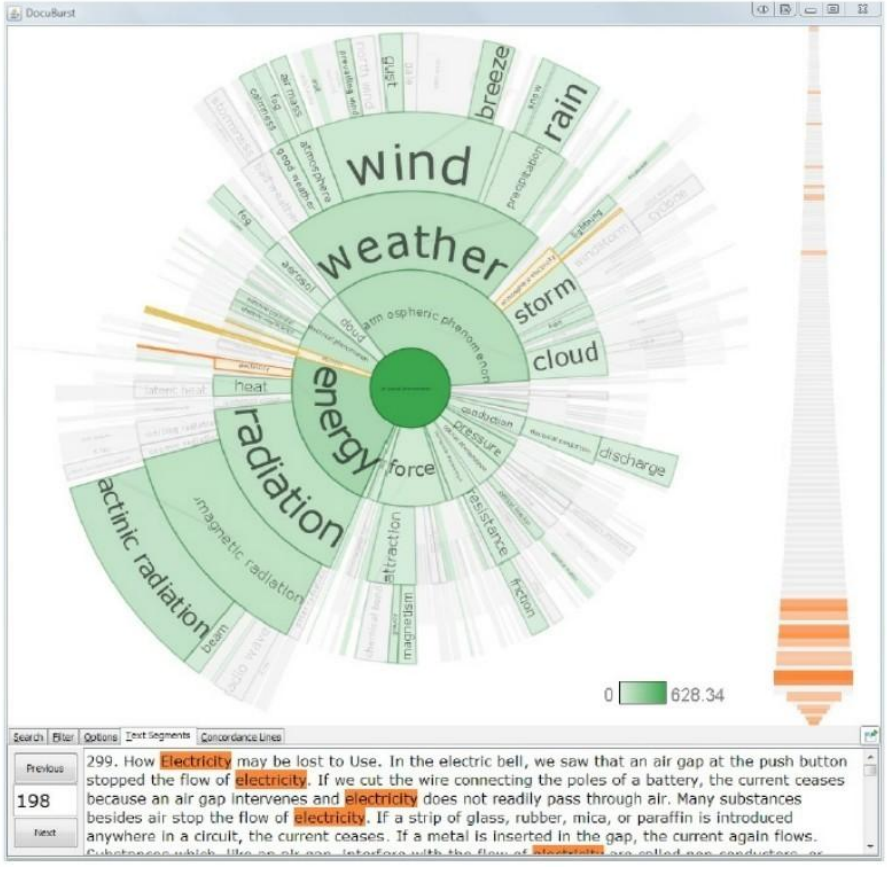
```
sort words by weight in descending order;
    FOR EACH word w:
        position of w= an initial point;
        WHILE w intersect with other words
            update position of w;
    END
END
```

EdWordle

- 允许用户移动和编辑单词，同时保留其他单词的邻域

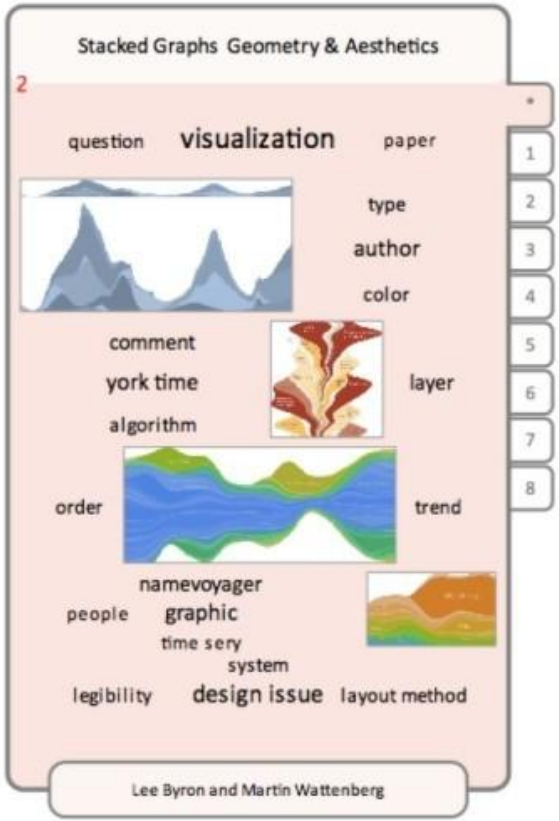
文档散(DocuBurst)

- DocuBurst采用径向布局，外圈的词汇是里圈词汇的下义词，圆心处的关键词是文章所涉及内容的最上层概述。每一个词的辐射范围覆盖其所有的下义词



文档卡片(Document Cards)

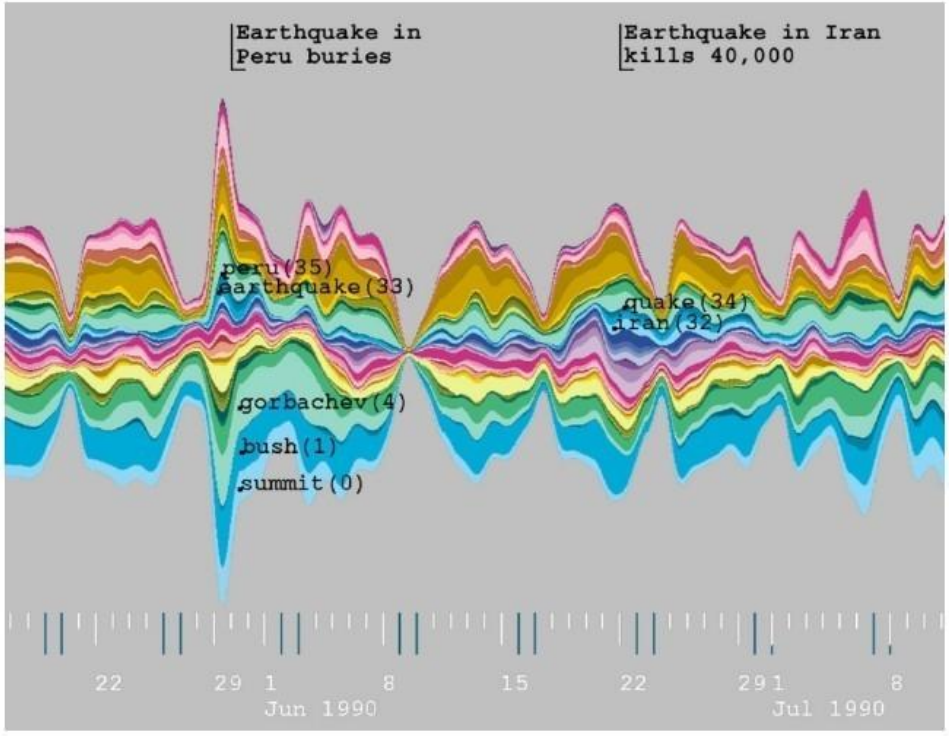
- 文档集合中每个文档的关键词和关键图片被紧凑地布局在一张卡片中，成为一张“扑克牌”，便于用户在不同尺寸的设备中查看和对比每个文档的信息



时序性的文本内容可视化

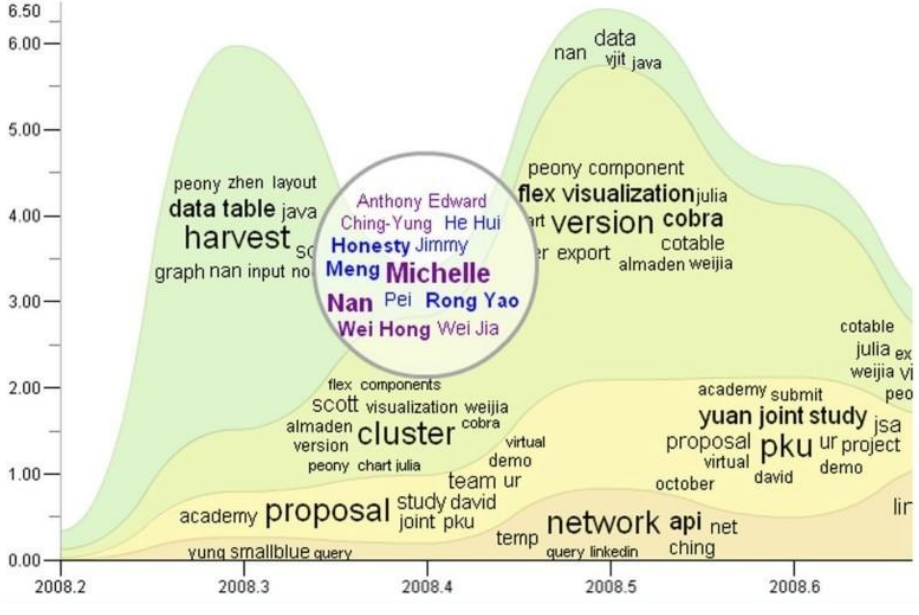
- 对于具有时间和顺序属性的文本，文本内容具有有序演化的特点
  - 一篇长篇小说有故事情节的发展变化
  - 新闻热点随着时间的推移而发生变化

- 主题河流(ThemeRiver)，采用河流作为可视原语来编码文档集合中的主题信息，将主题隐喻为时间上不断延续的河流



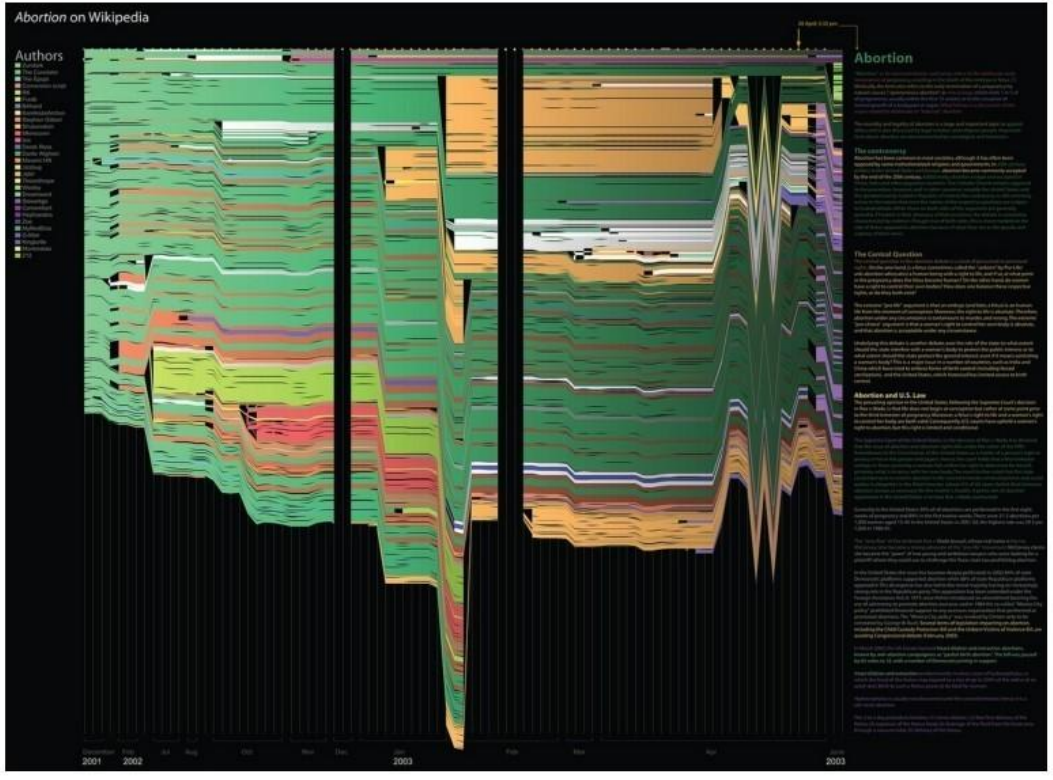
TIARA

- TIARA，将标签云技术与主题流结合，用其来描述文本主题在内容上随时间推进而发生的变化
- 此外，TIARA为每个文本主题在每个时间点上提取出不同的关键词，然后将这些词排布在相应色带上的相应位置，并用词的大小表示关键词在该时刻出现的频率
- 为了紧凑美观地排列主题支流，TIARA系统还设计了一系列自动调节支流顺序的算法



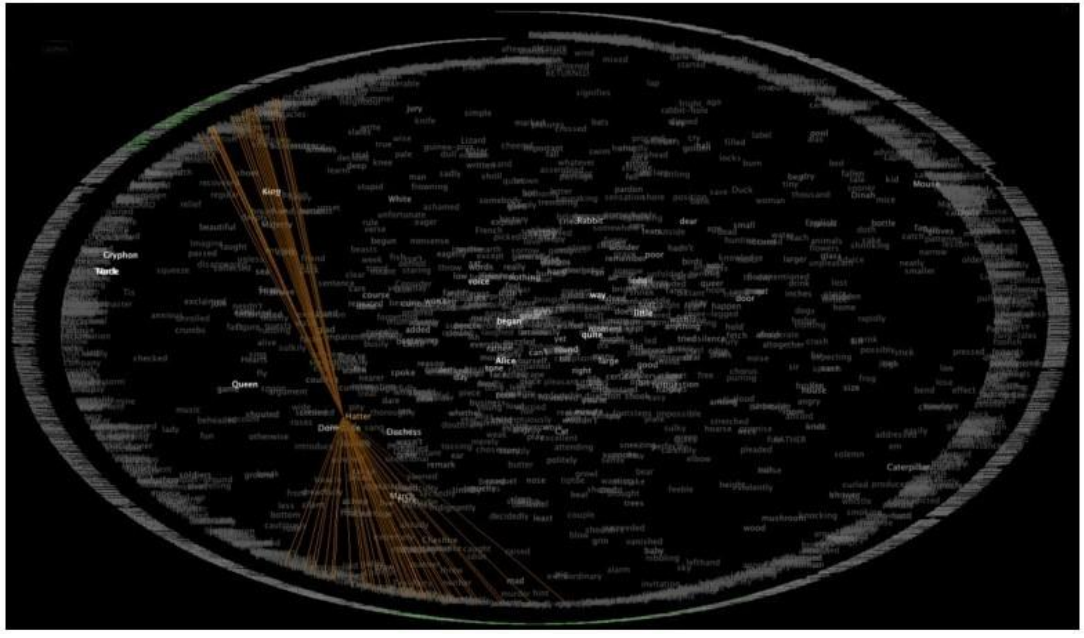
历史流(History Flow)

- 历史流(History Flow)，方法的设计初衷是可视地表达每个版本的维护者和他们所做的修改



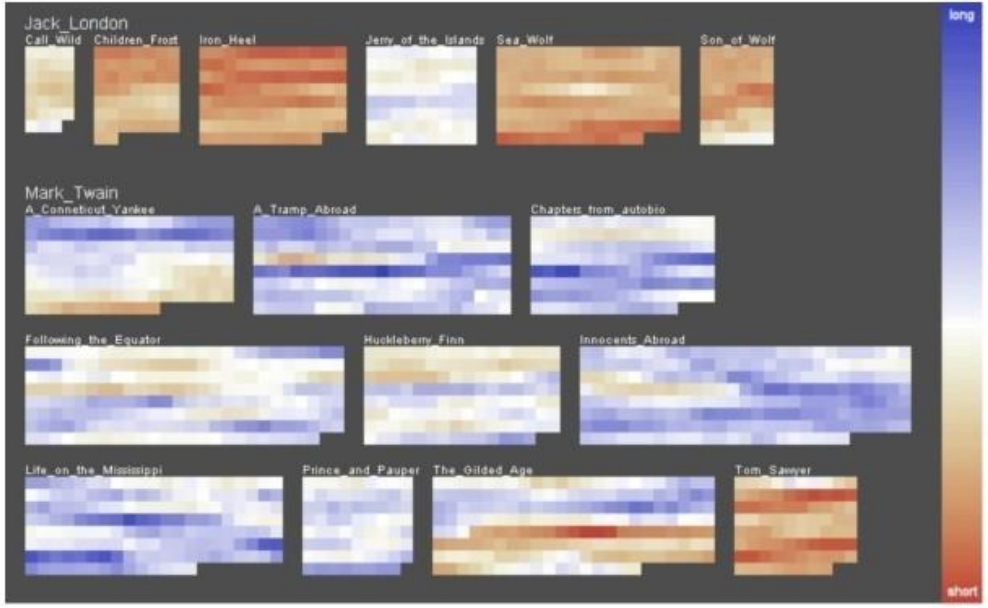
文本特征的分布模式可视化

- 文本弧(TextArc)

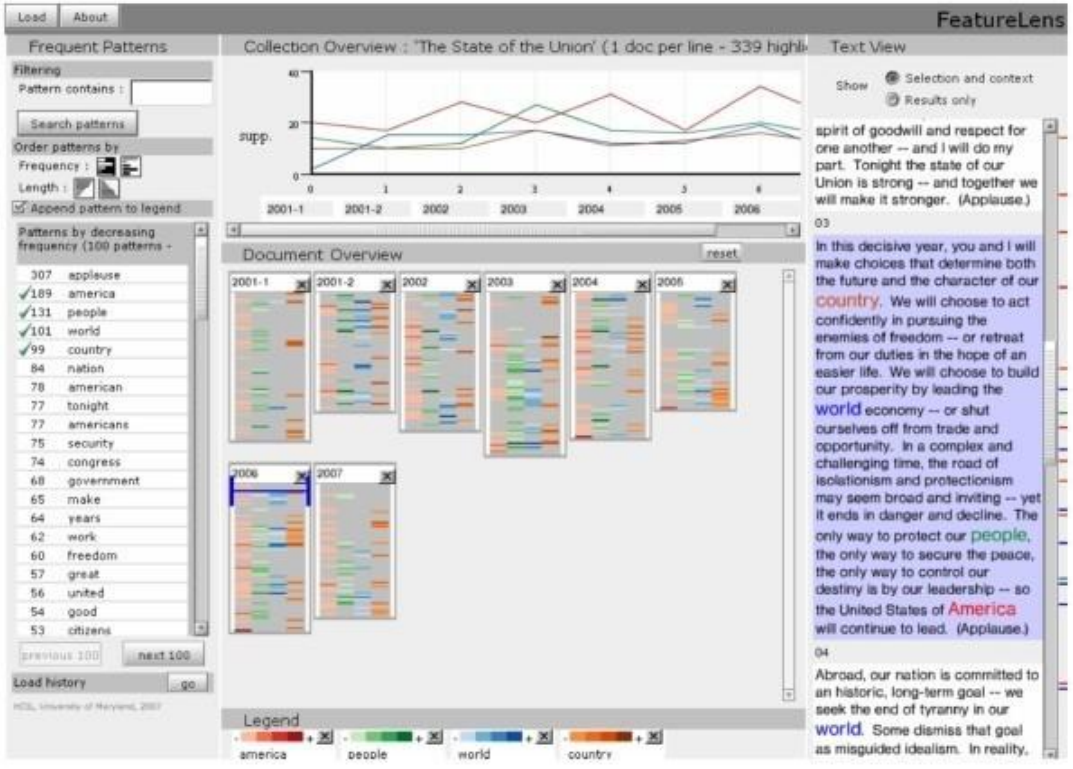


文献指纹(Literature Fingerprinting)

- 文献特征指纹，将特征在整个文本中的分布用一系列像素图(Pixel Chart)表达
  - 每一个像素块代表一段文本，一组像素块代表一本书

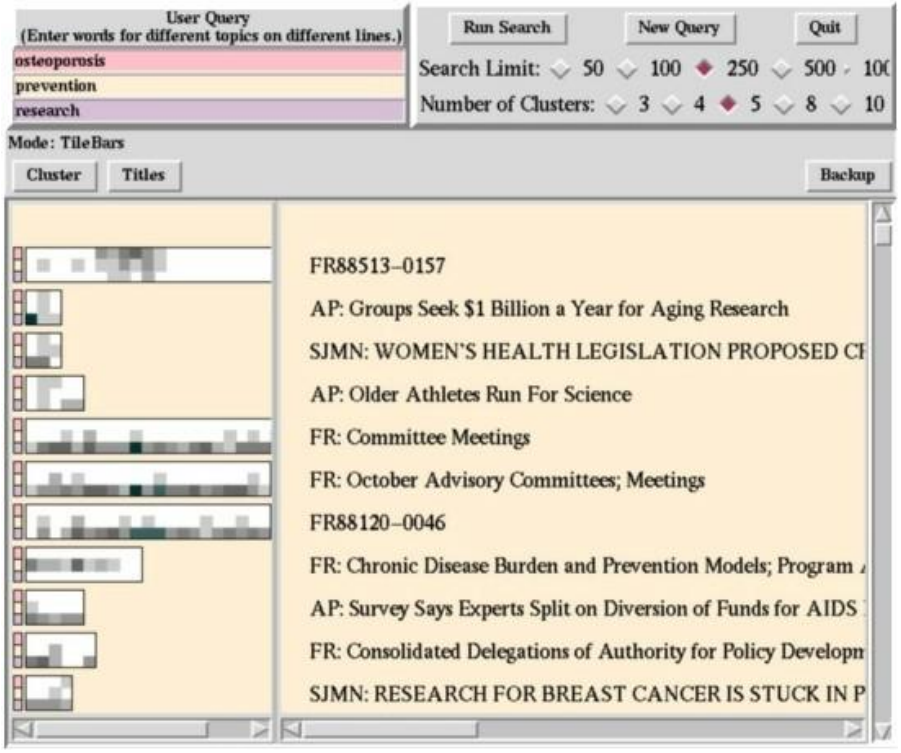


文本特征透镜(Feature Lens)



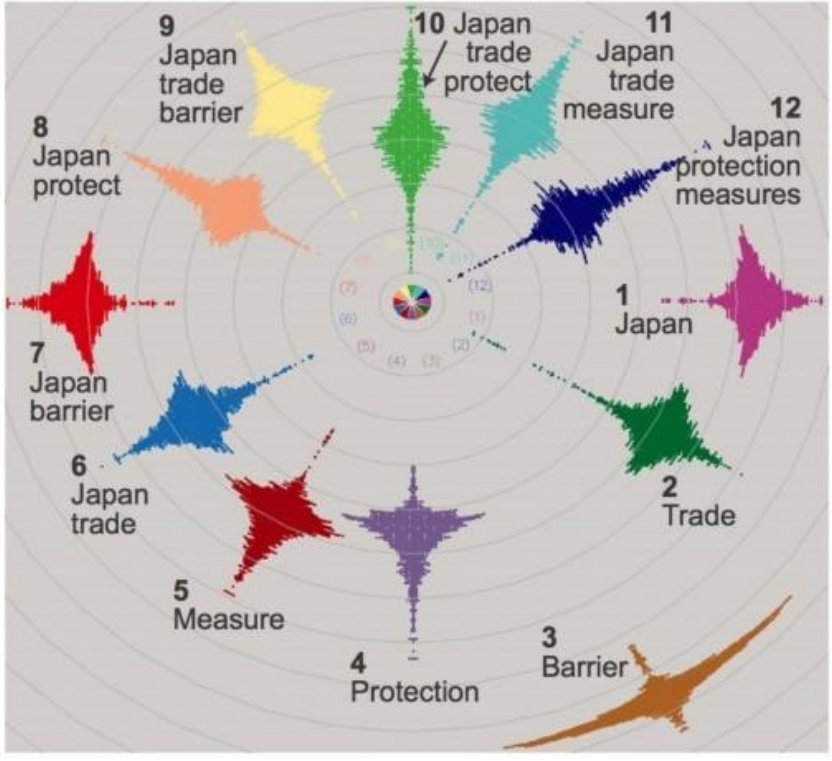
文档信息检索可视化

- TileBar
  - 帮助用户分析检索到的每篇文档和查询项间的匹配程度



Sparkler

- 通过可视化提供了查询项和检索到的文档集之间匹配程度的概览，也可用于比较不同查询项的检索结果之间的差异

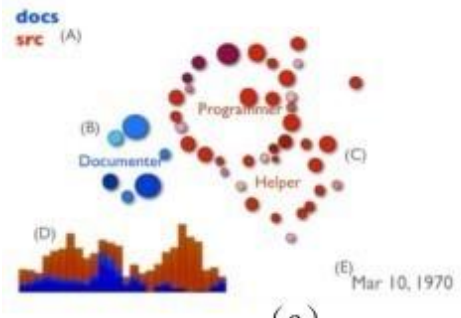


软件可视化

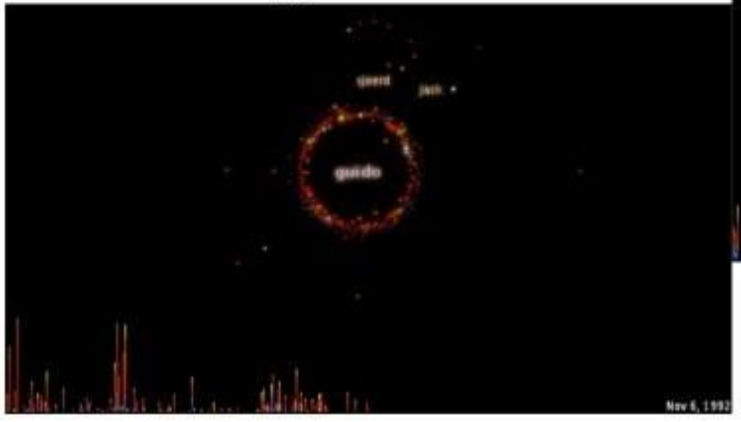
- SeeSoft[Eick1992]方法着眼于可视化每行代码的信息

Code_Swarm

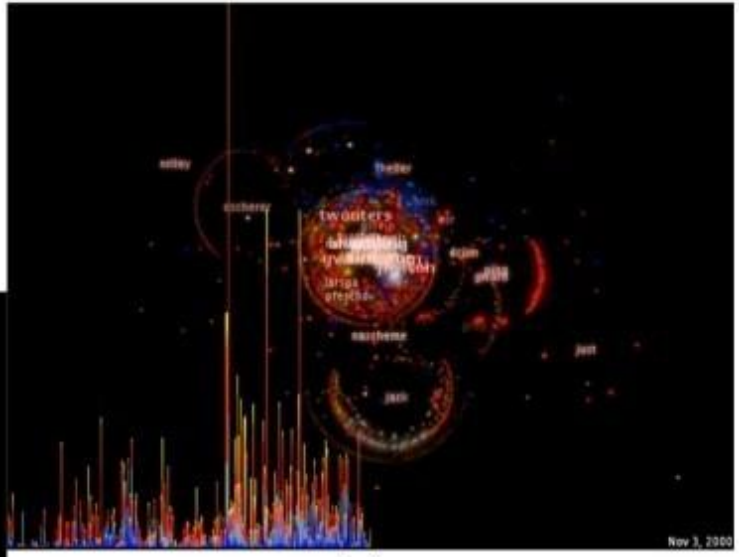
- 软件开发过程可视化
  - 开发和维护的文档类型、文档的提交时间、项目的参与人员



（a）



（b）

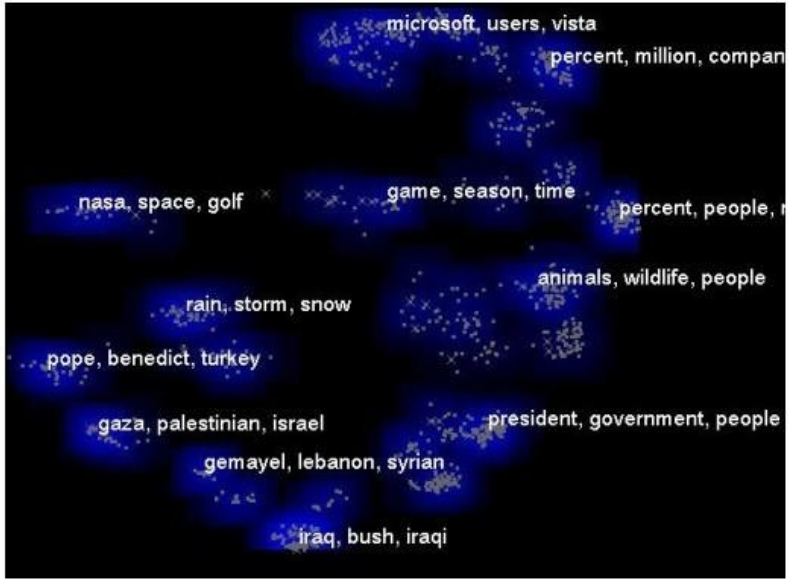


（c）

4 文本关系可视化

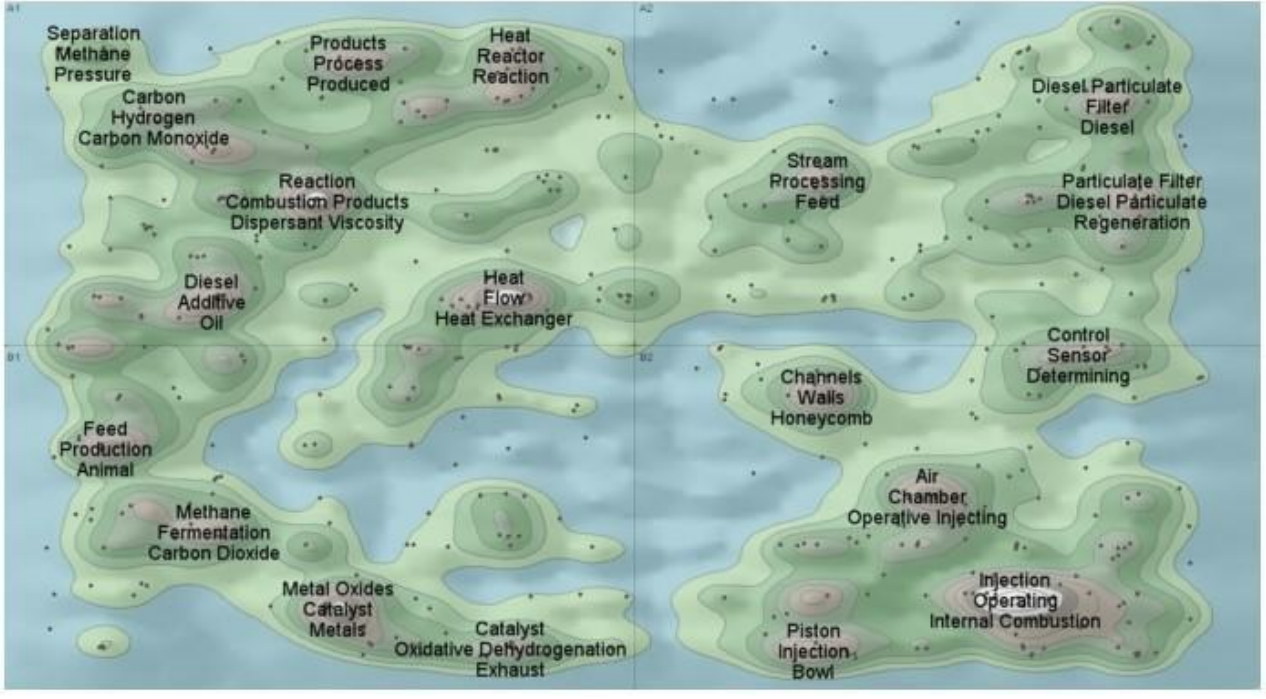
文档相似性可视化

- 对单个文档定义一个特征向量，利用向量空间模型计算文档间的相似性，并采用相应的投影技术呈现文档集合的关系
  - 主元分析(Principal Component Scaling,PCA)、多维尺度分析(Multidimensional Scaling,MDS)和自组织映射(Self-organizing Map,SOM)
- 星系视图(Galaxy View)



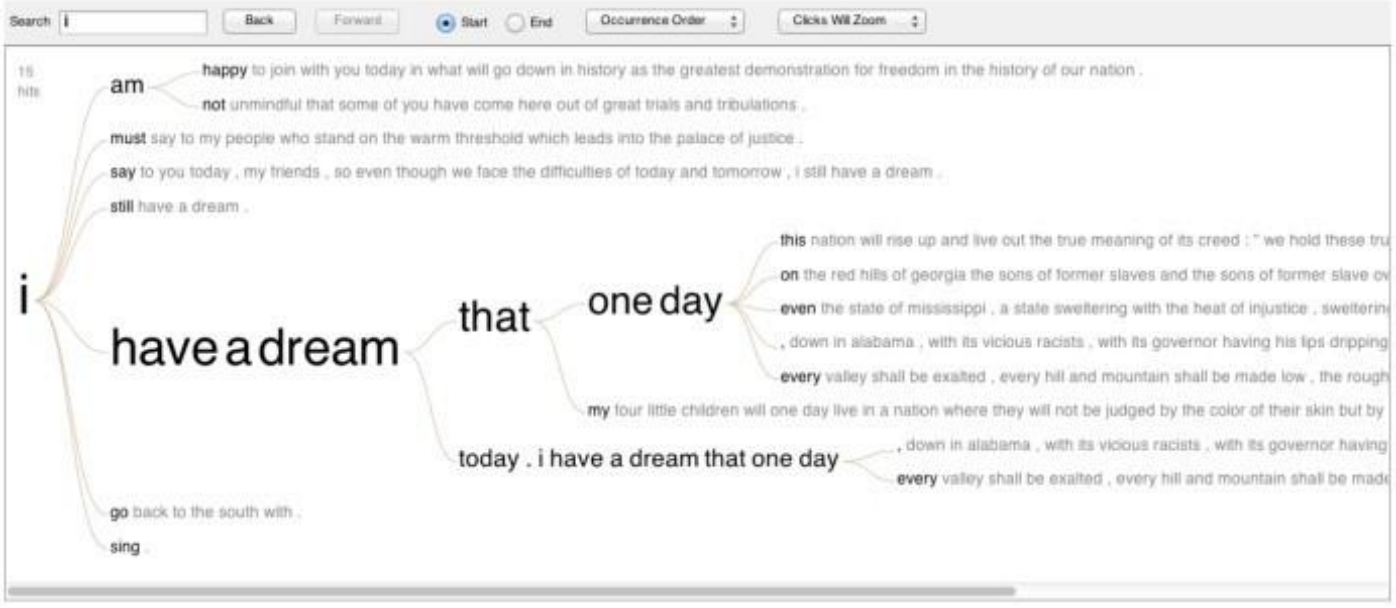
主题地貌(ThemeScape)

- 对星系视图方法改进：在其所计算的文档投影位置的基础上，采用等高线的方式可视表达文档集合中相似文档的分布情况



文本内容关联可视化

- 单词树(Word Tree)，从句法层面可视表达文本词汇的前缀关系

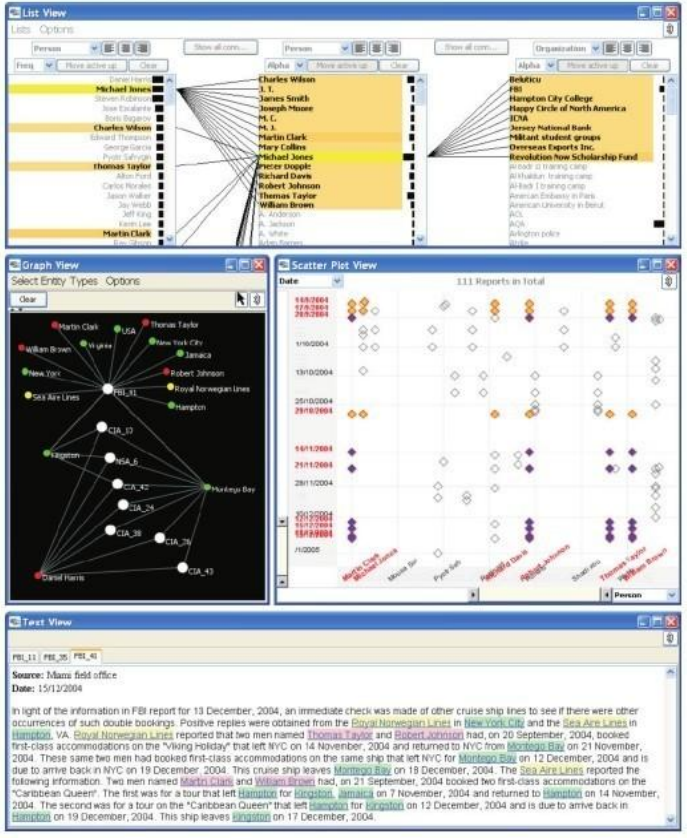


短语网络(Phrase Nets)、新闻地图(NewsMap)

- 短语网络(Phrase Nets)，采用节点-链接图展示无结构文本中语义单元彼此间的关系
- 树图方法也可用于刻画文本间的相似性

JigSaw

- 帮助用户分析集合中的文档以及文档中存在的实体间的相互关系



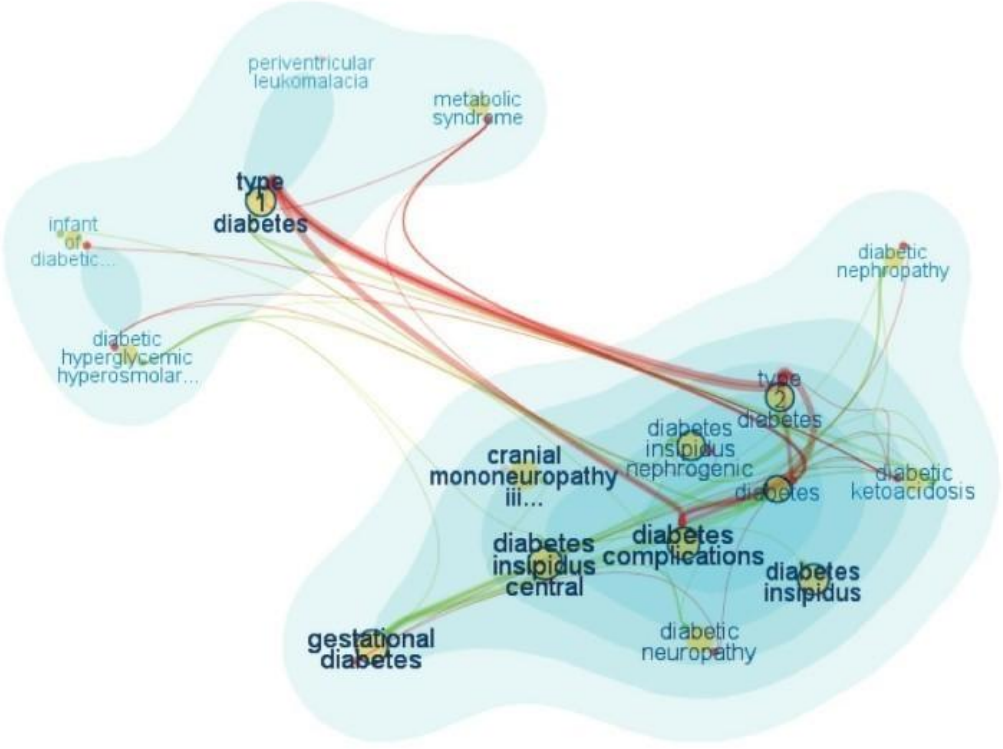
ContexTour

- 可视化文档集合所涉及的
  - 多个层面的内容，和各个层面间的关系

（a）Conference View　（b）Author View　（c）Keyword View

FacetAtlas

- 从文本信息的内容与关系的角度出发，分析并解释多层面的文本信息



5 文本情感分析可视化

情感分析

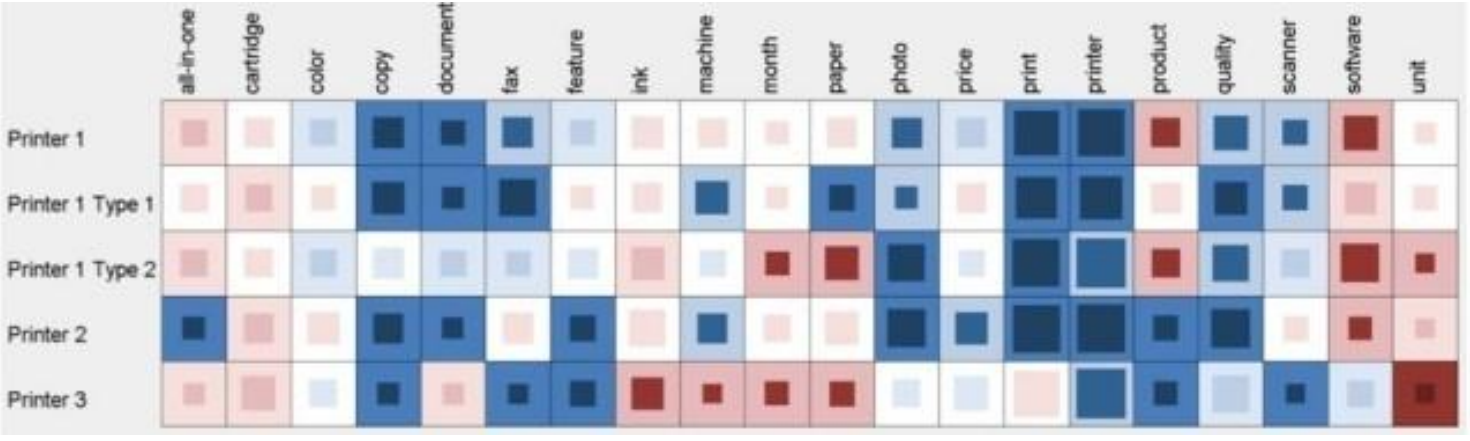
- 文本中往往蕴含了描述人类主观喜好、赞赏、感觉的情感信息

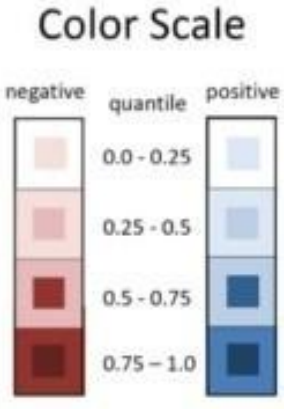
- 情感分析(又称意见挖掘，Sentiment Analysis或Opinion Analysis)，常被应用于论坛用户发言、社交网络、微博数据以及各种调研报告等文本中

- 情感分析的挖掘技术，可提取出文本中主观性的信息，并通常转化为一个区间分数
  - 一端为正面、积极的倾向
  - 另一端为负面、消极的倾向

顾客评价可视化 - 基于TFICF的用户反馈度量

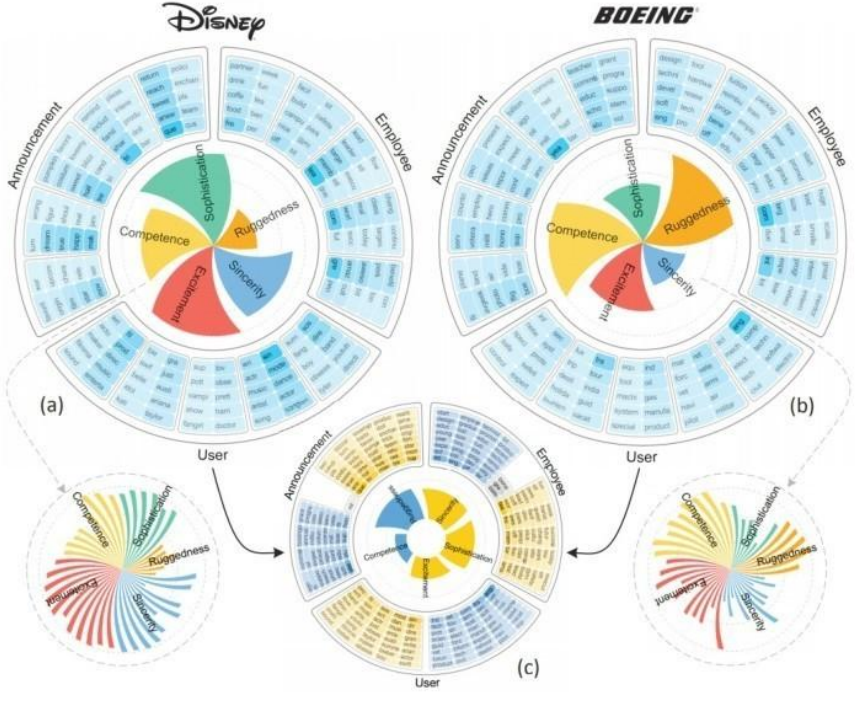
- 引入Term Frequency and Inverse Class Frequency(TFICF)，对打印机用户的反馈进行量化评分
  - 包括评价对象、评价属性和参与评价的用户人数三类信息
  - 矩阵行代表评价对象；列代表用户的评价属性；颜色传达用户的观念倾向程度；矩阵方格中的小格子表示参与评价的用户人数
  - 颜色代表用户对打印机的评价参数：红色代表消极，蓝色代表积极，透明度代表不同程度的评价





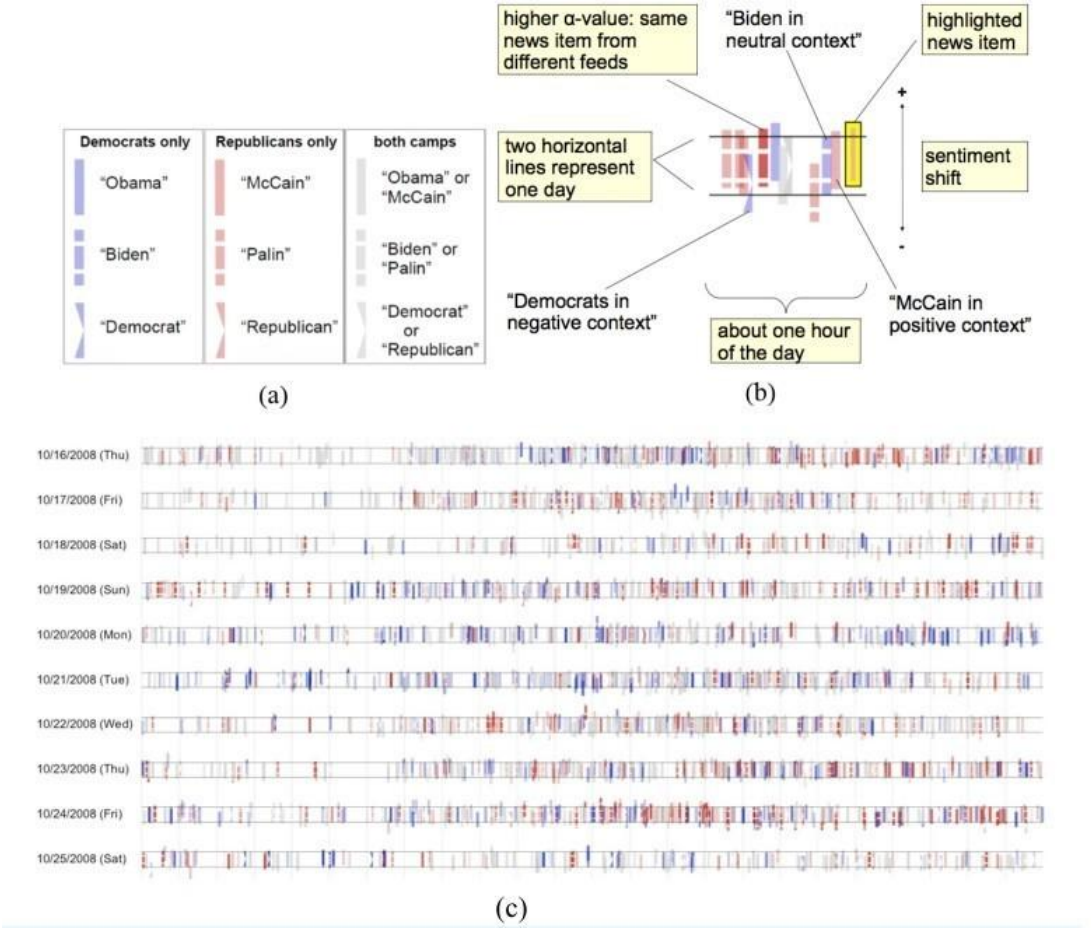
SocialBrands

- 综合利用了市场中的品牌个性框架与社会学计算方法，计算表现品牌个性的三要素：用户印象、员工印象、官方公告；并通过构建一个证据网络，解释品牌个性与驱使因素之间的关系



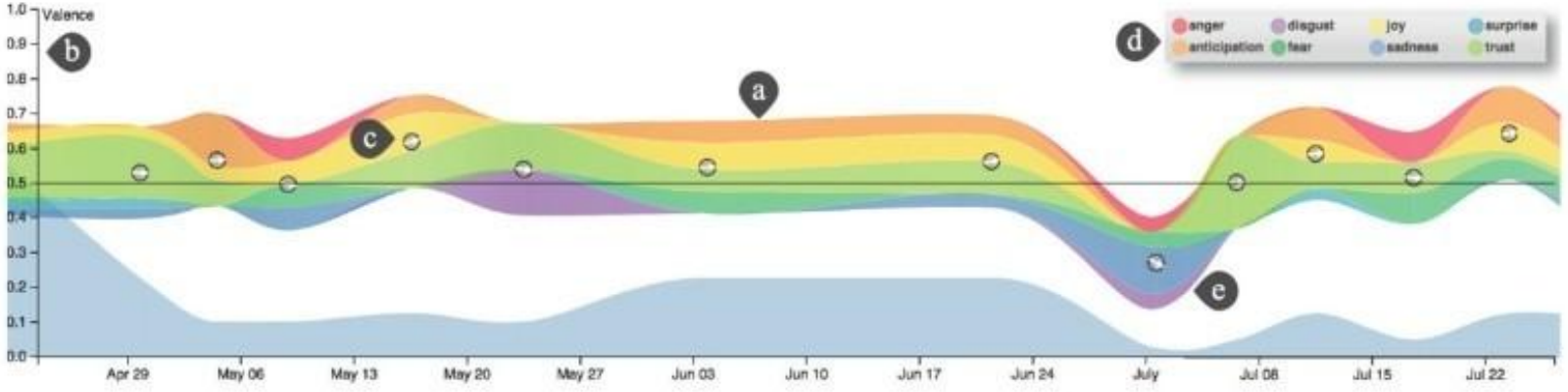
情感变化可视化

- 情感变化时序映射



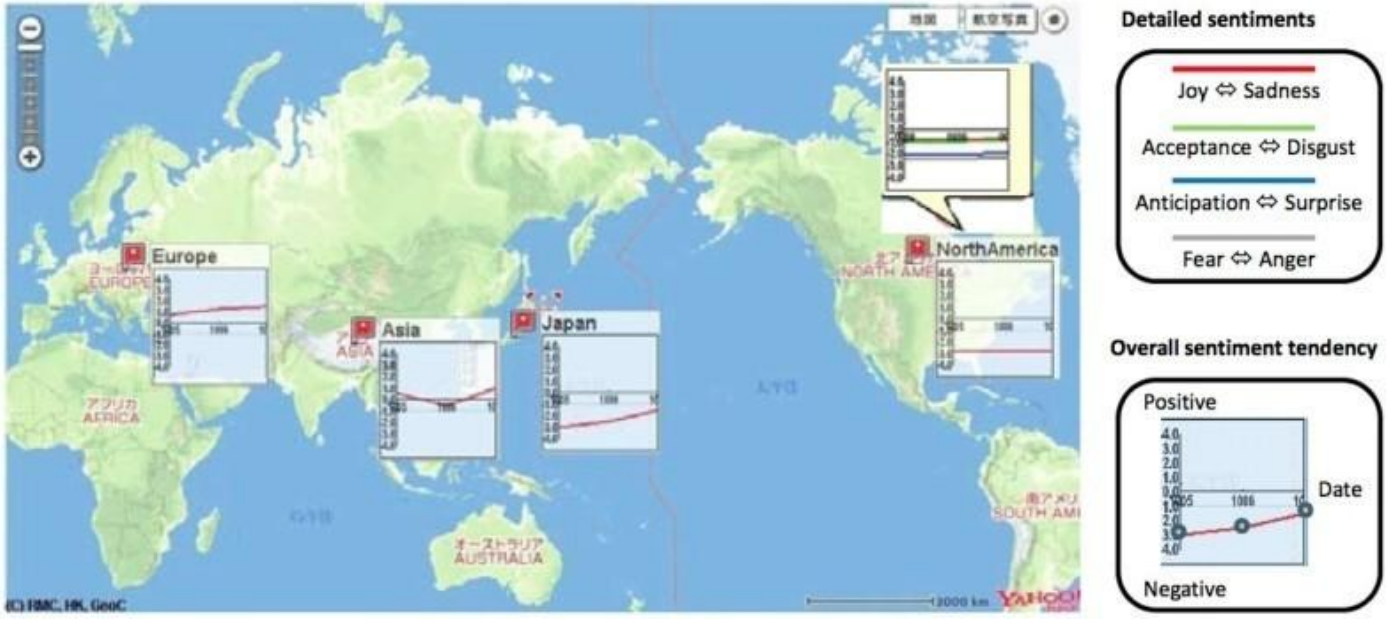
Pearl

- 自动检测某个个体在不同时间点表达的情绪，并基于聚类的方法对这些情绪进行总结，以揭示该个体的情感风格



情感差异可视化

- 情感地图：结合地理信息的新闻数据的情感分析可视化技术



6 总结

总结

- 文本可视化涉及文本信息提取技术、可视表达两个方面
- 文本可视化领域常用的文本可视化基础知识和方法以及文本信息提取技术，并从文本内容、文本关系、情感分析的角度阐述了文本可视化的研究内容和现有成果
- 文本信息没有空间位置等结构化信息。如何将没有空间结构属性的文本信息转换为用户乐于接受的二维或三维空间的可视表达结果是文本可视化面临的一个核心问题

- 在未来的文本可视化研究中，如何将文本分析模型和信息可视化技术无缝结合，如何更好地处理海量、时变、具备多重语义的文本信息是极大的研究挑战