虚拟场景中社会信息的组织和显示

殷 钊[1,2], 王 衡[1,2], 汪国平[1,2]

[1] (北京大学图形与交互技术实验室 北京 100871)
[2] (北京市虚拟现实与可视化工程技术研究中心 北京 100871)
(yinzhao@graphics.pku.edu.cn)

摘 要: 为提高用户对于网络中海量社会信息的浏览效率, 减轻交互负担, 提出一种多线索的社会信息组织方式, 对于其中的主题线索, 提出一种基于 LDA 的主题自动提取方法, 以地点为单位聚合大量消息来挖掘主题, 然后根据生成概率模型为每条消息选择主题; 基于多线索的信息组织, 设计并实现一种在虚拟场景中显示社会信息的可视化方法, 采用经典的可视化结构和直观的交互操作. 实验结果证明了主题提取方法的有效性, 多线索的信息组织和可视化方法能够有效地提高交互效率, 可满足用户浏览信息的不同需求.

关键词: 社会信息; 主题模型; 信息可视化; 用户界面
中图法分类号: TP391

Organizing and Visualizing Social Information in Virtual Scenes

Yin Zhao[1,2], Wang Heng[1,2], and Wang Guoping[1,2]

[1] (*Graphics and Interaction Lab, Peking University, Beijing* 100871)
[2] (*Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Beijing* 100871)

Abstract: To improve interaction efficiency and reduce interactive burden for browsing massive social information on the Internet, a multi-cue information organizing method and a LDA-based topic extraction approach are proposed. A large amount of messages of a location are aggregated to discover the topics, and then the topic of each message is selected using generative probabilistic model. A visualization method is designed and implemented to display social information in virtual scenes based on multi-cue information organizing, which utilizes classical visual structures and natural interaction. Experiment results prove the effectiveness of the topic extracting method. Interaction efficiency is improved effectively and different requirements for browsing information are met with the multi-cue information organizing and visualization method.

Key words: social information; topic modeling; information visualization; user interface

近年来, 各种社交网络在国内外迅速崛起, 如 Facebook、Twitter、新浪微博、人人网等, 吸引了越来越多的用户在其上发布、交流和分享信息. 基于位置的社交网络(location-based social networks, LBSN)是一种特殊的社交网络, 人们发布的信息不仅具有传统的信息属性, 同时可以共享自己的地理位置. 随着智能手机的普及和移动互联网的崛起, 人们在发布信息时共享自己的地理位置

收稿日期：2014-08-10; 修回日期：2014-09-28. 基金项目：国家“九七三”重点基础研究发展计划项目(2010CB328002); 国家科技支撑项目(2013BAK03B07); 国家自然科学基金项目(61232014, 61121002, 61173080). 殷 钊(1989—), 男, 硕士, 主要研究方向为人机交互、虚拟现实; 王 衡(1960—), 女, 博士, 副教授, 硕士生导师, 主要研究方向为人机交互、图像处理、计算机图形学等; 汪国平(1964—), 男, 博士, 教授, 博士生导师, 主要研究方向为计算机图形学、人机交互、虚拟现实等.

十分简单易行. 位置是人日常生活中最重要的方面之一, 有了位置的参与, 社交网络从一个虚拟世界变得更加贴近现实世界, 用户可以在日常生活中发布自己在不同地点的见闻、心情状态, 其中的社会信息也由于更加真实而具有更大的利用价值.

网络中汇聚了公共的或个人发布的文字、图片、视频等海量社会信息, 通过浏览这些信息, 人们可以得到自己关注和感兴趣的信息, 从而满足一定的生活需求. 然而, 社会信息的极度膨胀及其本身固有的多样性使得用户在寻找想要的信息的过程中要付出大量的时间和精力.

从提高用户的浏览效率和减轻用户的交互负担的目的来讲, 对于社会信息的处理应当着重研究信息的组织和信息的呈现 2 个方面. 社会信息具有时间、地点、用户等属性, 可以基于这些属性根据多种线索对信息进行组织, 用户能够通过选择不同的线索快速获得感兴趣的信息. 在得到有效的组织方式之后, 还需要一种合理的可视化方式将信息以直观的、有意义的形式呈现给用户, 并通过良好的交互方式提供多种线索的选择, 使用户获得最好的交互体验.

本文主要以基于地理位置的社会信息为对象, 讨论了信息的组织方式和在虚拟场景中的可视化显示, 帮助用户更加方便、迅速地浏览和查找所需信息, 提供便捷的工作生活服务.

1 相关研究

许多研究者提出了一些有效的线索来组织社交网络中的社会信息, 从而能使用户只关注自己感兴趣的信息.

社交网络包含了大量的用户发布的关于事件、新闻的信息, 在一定程度上具有新闻媒体的特性[1]. 通过“新闻”对大量的信息重新组织和过滤, 能够向人们提供最有价值的信息. Marcus 等[2]提出了 TwitInfo 系统, 能够根据 Twitter 上用户发布的信息实时地可视化和汇总其中发生的事件, 用户可以查看某个事件相关的 tweets. Nichols 等[3]采用一种无监督学习算法, 挖掘 Twitter 中与体育比赛相关的 tweets, 将其自动汇聚为一场体育比赛的新闻报道. 人们日常生活的事件具有更加普遍意义的新闻价值, Zuo 等[4]研究了如何在新浪微博中筛选具有“新闻”价值的信息.

通过“主题”对社会信息进行重新组织也是一种有效的方式, 用户能够根据自己的兴趣按照主题进行更有目的的浏览. Chen 等[5]做了丰富的实验, 验证了在 Twitter 中通过主题相关性向用户推荐信息的有效性. Esparza 等[6]设计了 CatStream 系统, 将 Twitter 中的信息按照主题进行过滤, 用户可以只专注于自己感兴趣的主题相关的信息. Bernstein 等[7]提出了一种挖掘 tweets 主题的方法, 并设计了一个 Twitter 客户端 Eddi, 将用户关注的 tweets 整理为不同的主题, 使用户能够根据兴趣浏览信息, 并借助搜索引擎来得到一条 tweet 的主题, 对于单个主题能够达到 40%的准确率.

另一方面, 良好的可视化方法对于显示社会信息具有重要的意义. Ren 等[8]提出了一个可视化的分析系统 WeiboEvents, 以转发树结构可视化展示并分析新浪微博中的转发事件, 用户可以基于关键词对转发树进行过滤和细化. Tang 等[9]提出了 LifeCircle, 能够可视化地聚合显示一个新浪微博用户很长时间段内发布的微博概览, 基于一个圆圈结构融合了多种线索进行信息显示, 借助 LifeCircle, 用户可以更好地回忆往事.

对于基于地理位置的社会信息, 地图是最经典的用于体现地理位置的工具. Gao 等[10]提出了一个基于地图的新闻可视化系统 NewsView, 当用户浏览一篇新闻文章时, 能够自动地生成一个交互式的带注解的地图, 用以展示该新闻相关的地点和话题下的其他新闻. MacEachren 等[11]提出了一个支持态势感知的基于地理位置的可视化分析系统 SensePlace2, 基于地图呈现 Twitter 中具有显式或隐式地理位置信息的 tweets, 能够展示出地点、时间、主题等多种线索.

本文提出了多种组织社会信息的线索, 并研究了一种适用于基于位置的社会信息的主题抽取方法, 设计并实现了一种在基于地理位置的虚拟场景中直观合理地显示社会信息的新颖的可视化方法, 能够大大提高用户浏览信息的效率, 减轻用户的交互负担.

2 多线索的信息组织和主题提取

本文引入“线索”的概念, 把用于组织社会信息所基于的一个侧面称为一个线索, 提出了基于位置的社会信息中的 4 种线索: 地点、用户、时间、主题. 基于多种线索, 可以有效地梳理相关联的社会信息, 用户能够按照自己的需求选择相关线索

来浏览感兴趣的信息.

地点是位置社会信息中最基本也是最重要的线索，将大量的信息按照不同地点分别组织是最直观的方式;用户是社会信息的重要组成部分，不同用户感兴趣的信息所属的用户和社交圈具有很大差异，可以按照不同用户或不同用户组来组织信息；时间是一个不可或缺的附加属性，信息随着时间的变化具有不同的价值，根据时间浏览消息是一个广泛的需求，一定时间段内信息的时间序排列是常用的信息组织方法；主题则是一个语义层面的概念，用户发布的消息中蕴含着语义含义，主题可以先导地降低用户对消息语义的认知负担，基于主题对社会信息进行分类组织，能够体现出信息之间的语义关联，使用户对大量的信息有直观、简单的认知了解. 不同的线索之间并不是互斥的，同时指定多种线索，则信息按照这些线索进行组织，用户可以只查看与当前线索相关的社会信息. 通过这 4 种线索能够极大地滤过信息，大大缩减相关社会信息的范围.

主题是蕴含在消息内容中的语义信息，并不能直接得到，需要通过某些方法从消息内容中获取. 在基于位置的社会信息中，用户在地点上发布的消息有许多会与这个地点有语义相关性，因而相比于一般的社会信息具有更明显的特征. 本文充分利用其特有的特征，基于隐式狄利克雷分配(latent Dirichlet allocation, LDA)模型对地点上的消息进行建模，构建每个地点的主题模型，然后利用地点的主题模型确定其中的消息的主题.

2.1　LDA

LDA 是一个基于概率的生成模型，能够用来挖掘文章集合中潜在的主题信息. 它基于一个常识性的假设：语料库中的所有文档均共享一组相同的隐含主题[12].

假设语料库中有 M 文档，M 个文档包含 K 个潜在的主题，词汇表中共有 V 个单词. 对于语料库的每个文档 d, LDA 定义了如下的生成过程[13]:

Step1. 选择文档 d 的单词总数 N, N ~Possion(ξ).

Step2. 生成文档 d 的主题分布$\boldsymbol{\theta}$, $\boldsymbol{\theta}$~Dir(α).

Step3. 对于文档 d 的每个单词：

Step3.1. 从 d 的主题分布$\boldsymbol{\theta}$ 中选择一个主题 z.

Step3.2. 从主题 z 的单词分布$\boldsymbol{\varphi}$中选择一个单词 w.

上述的生成过程涉及到 2 个概率分布：每个文档到 K 个主题的概率分布，记为$\boldsymbol{\theta}$; 每个主题到 V 个单词的概率分布，记为$\boldsymbol{\varphi}$. 概率分布$\boldsymbol{\theta}$ 有一个带有超参数 α 的 Dirichlet 先验分布，概率分布$\boldsymbol{\varphi}$有一个带有超参数 β 的 Dirichlet 先验分布，即 $\boldsymbol{\theta}$~ Dir(α), $\boldsymbol{\varphi}$~Dir(β).

LDA 的目标是对文档集合进行建模，根据已有的数据推断文档的生成过程，重建$\boldsymbol{\theta}$和$\boldsymbol{\varphi}$ 从而挖掘出文档中潜在的主题信息. 因此，在离散的数据上, LDA 模型最终可以得到 2 个矩阵：文档—主题的概率分布矩阵$\boldsymbol{\theta}$和主题—单词的概率分布矩阵$\boldsymbol{\varphi}$.

2.2　位置社会信息的 LDA 建模

本文以地点为单位，聚合一个地点上用户在最近一段时间发布的大量消息，将其建模为 LDA 模型中的一个文档. 通过 LDA 方法构建每个地点的主题模型，然后可以利用该模型为地点上的该时间段内的每条消息选择合理的主题. 综合来说，一个地点上大量消息的集合对应 LDA 模型中的文档层，消息文本中的单词对应 LDA 模型中的单词层，待抽取的地点主题对应 LDA 模型中的主题层.

为了将地点上的消息聚合为一个文档，每条消息需要进行文本的预处理:

Step 1. 利用中文分词工具进行分词.

Step 2. 去除标点符号和长度为 1 的词.

Step 3. 根据词性进行过滤.

主要过滤掉数量词和 URL 链接，它们一般并不能直接表达出语义含义.

Step 4. 停用词过滤.

停用词(stop words)指的是文章中出现频率太高，却没有太大实际意义的词，主要是一些副词、虚词和语气词等，如“的”“是”“太”等.

对一个地点的所有消息均进行预处理之后，便聚合成为了一个文档，该文档实际是一个单词集合. 对所有地点都处理完毕之后，即得到了所有的文档. 利用 LDA 方法进行分析，可以得到文档—主题分布$\boldsymbol{\theta}$ 和主题—单词分布$\boldsymbol{\varphi}$, 也就得到了每个地点 m 的主题模型 $\boldsymbol{\theta}_m = p(z \mid d = m)$.

2.3　主题词提取

得到每个地点的主题模型之后，即可利用该模型为地点上的消息提取主题，本文通过如下 2 个步骤得到每条消息的主题词：

Step 1. 生成每个地点的候选主题集合.

地点的主题模型是一个概率分布，并不是其中的所有主题都与该地点有较强的语义关联，首先应当确定一个符合该地点语义的候选主题集合.

每个地点的主题应当在一定程度上体现地点的特性，尽量避免过于泛化的主题. 本文基于经典的 TF-IDF(term frequency-inverse document frequency)思想对主题在地点中的重要性进行重排序. 对于一个主题 z_j 个地点 d_i, z_j 在 d_i 中的重要性权重通过

$$T_{i,j}=\theta_{i,j}\times\left(\theta_{i,j}\Big/\sum_{k=1}^{M}\theta_{k,j}\right)$$

计算. 其中, $\theta_{i,j}$ 表示主题 z_j 在地点 d_i 上的概率; M 表示地点总数.

对于一个地点经过重要性权重重排序之后的主题列表, 需要再从中选择若干个语义含义较强的主题作为候选主题. 首先设定一个概率阈值 p_t, 若一个主题下概率最高的 3 个单词的平均概率大于该阈值, 则认为该主题语义较强. 最终, 为每个地点选择权重最高且语义较强的前 K_c 个主题作为候选主题.

还存在一个问题, LDA 只是将主题以词汇表中单词的概率分布来表示, 然而抽象的主题需要被解释为一个主题词才能够符合用户的认知并易于用户理解. 本文同样采用 TF-IDF 思想选择一个主题的单词分布中重要性权重最大的单词作为该主题的主题词, 如

$$W_{i,j}=\varphi_{i,j}\times\left(\varphi_{i,j}\Big/\sum_{k=1}^{K}\varphi_{k,j}\right).$$

对于一个单词 w_j 和一个主题 z_i, $\varphi_{i,j}$ 表示单词 w_j 在主题 z_i 上的概率, K 表示主题数.

本文只将主题词的选取范围定位在一个主题中概率最高的前 10 个单词内, 对每个单词计算上述权重, 因为通常前 10 个单词已经足以表征主题的语义.

Step 2. 为每条消息选择主题.

LDA 定义了一个文档中每个单词的生成过程, 得到一个地点的候选主题集合之后, 对于一条消息, 本文通过模拟其中每个单词的生成计算它们的生成概率, 最后确定该消息的主题. 由于一般消息的内容较短, 本文仅为其选择一个主题.

确定一条消息的主题的步骤如下：

Step1. 进行与 2.2 中相同的预处理过程, 得到其单词集合.

Step2. 对于每个单词 w_i, 通过

$$p(w_i\mid z_k)=p(z_k\mid d=d_m)\cdot p(w_i\mid z=z_k)=\theta_{m,k}\cdot\varphi_{k,i}$$

计算其所属地点 d_m 的候选主题集合中每个主题 z_k 对该单词的生成概率. 其中, $p(z_k\mid d=d_m)$ 和 $p(w_i\mid z=z_k)$ 分别对应文档—主题分布 $\boldsymbol{\theta}$ 和主题—单词分布 $\boldsymbol{\varphi}$ 中的元素, 表示通过文档 d_m 选择主 z_k 题的概率和主题 z_k 选择单词 w_i 的概率.

Step3. 对于每个单词 w_i, 得到候选主题集合中使其取得最大生成概率的主题, 即

$$z_{\max}(w_i)=\arg\max_{z}p(w_i\mid z).$$

Step4. 对于该消息, 统计所有单词的最大生成概率主题 $z_{\max}(w_i)$的出现次数, 取出现次数最多的主题作为该消息的主题, 并以相应的主题词表示.

通过上述方法, 可以得到每条消息的主题及主题词. 特别地, 由于对一条消息的文本进行预处理之后所得的单词集合可能为空, 此时上述方法无法使用. 这些消息一般是不包含语义含义的, 直接为其赋予主题词“其他”.

3 社会信息的可视化显示

本文基于多线索的信息组织方式,设计并实现了一种在基于位置的虚拟场景中合理地显示社会信息的可视化方式, 其实现基于北京大学图形与交互技术实验室研发的虚拟现实平台 ViWo. 该可视化方式主要包括以地点为主线索的信息显示、多线索的信息显示、基于位置的信息解聚显示等特点.

3.1 以地点为主线索的信息显示

用户在虚拟地球上漫游时, 地理位置是最基本也是最受关注的信息, 因此地点线索是呈现社会信息的最主要线索. 本文采用兴趣点的概念, 用户在地理位置上发布消息时首先选择此地理位置上的一个兴趣点, 然后将消息发布至该兴趣点上. 如图 1 所示, 兴趣点采用最经典的图标结构标志在其对应的地理位置上, 并细分为不同的类型, 如餐馆、学校、公交车站等; 不同类型的兴趣点采用不同的图标标志, 用户通过侧边栏选择要查看的兴趣点类型. 点击一个兴趣点图标, 与该地点相关的信息汇聚显示在一个弹出的气泡内, 气泡也是位置社会信息可视化中经典的可视化结构.

3.2 多线索的信息显示

除了以地点为主线索的信息呈现, 本文在地点的气泡页面中提供用户、时间、主题等线索的选择, 用户可以通过直观的交互方式根据需求选择线索, 从而逐步定位到感兴趣的相关信息进行浏览.

地点的气泡页面详细结构如图 2 所示. 其中 a 部分用于提供线索的选择和显示, 主要为时间和用户线索. “2014-02-01 到 2014-05-10”表示选择了该时间段, 在未选择时间时, 该按钮显示白色的“时间”二字, 提供“今天”、“最近一周”、“自定义范围”等选择. “用户”与其类似, 初始为白色的“用户”二字, 当选择了用户线索后, 高亮显示

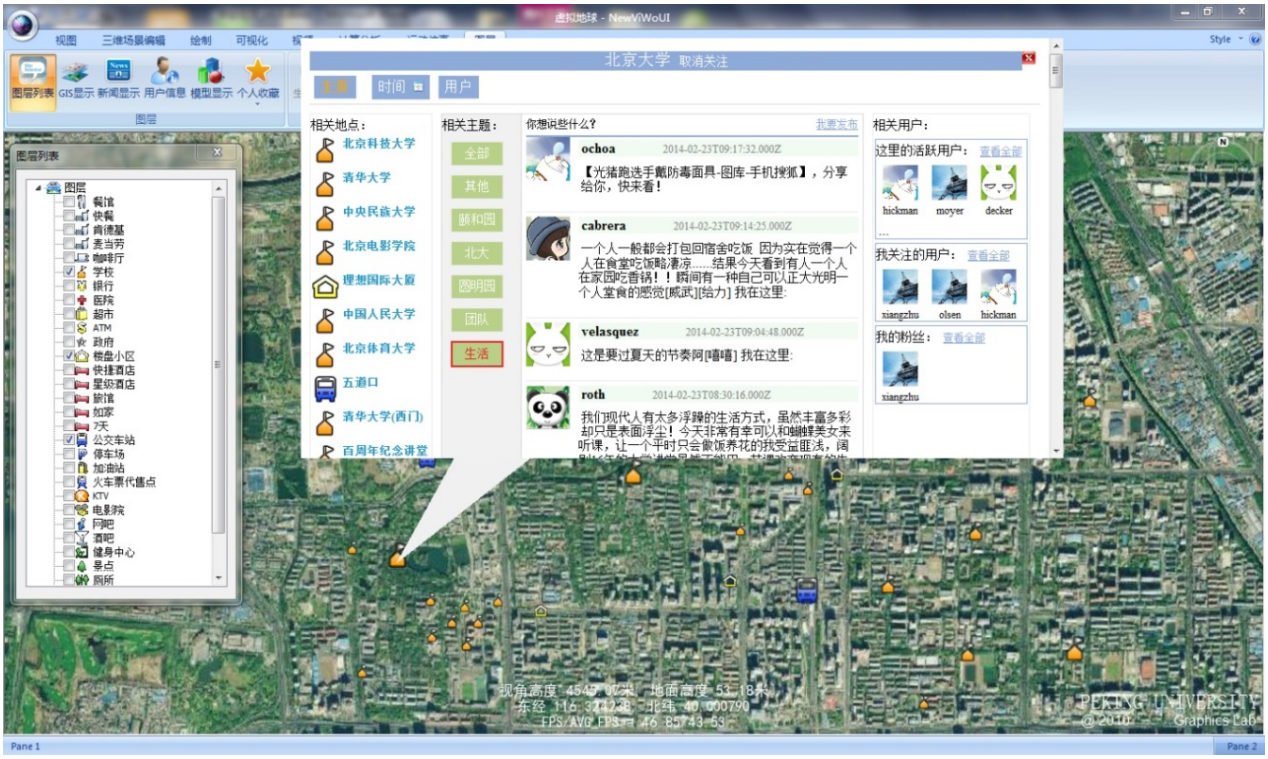

图 1　以地点为主线索的信息显示

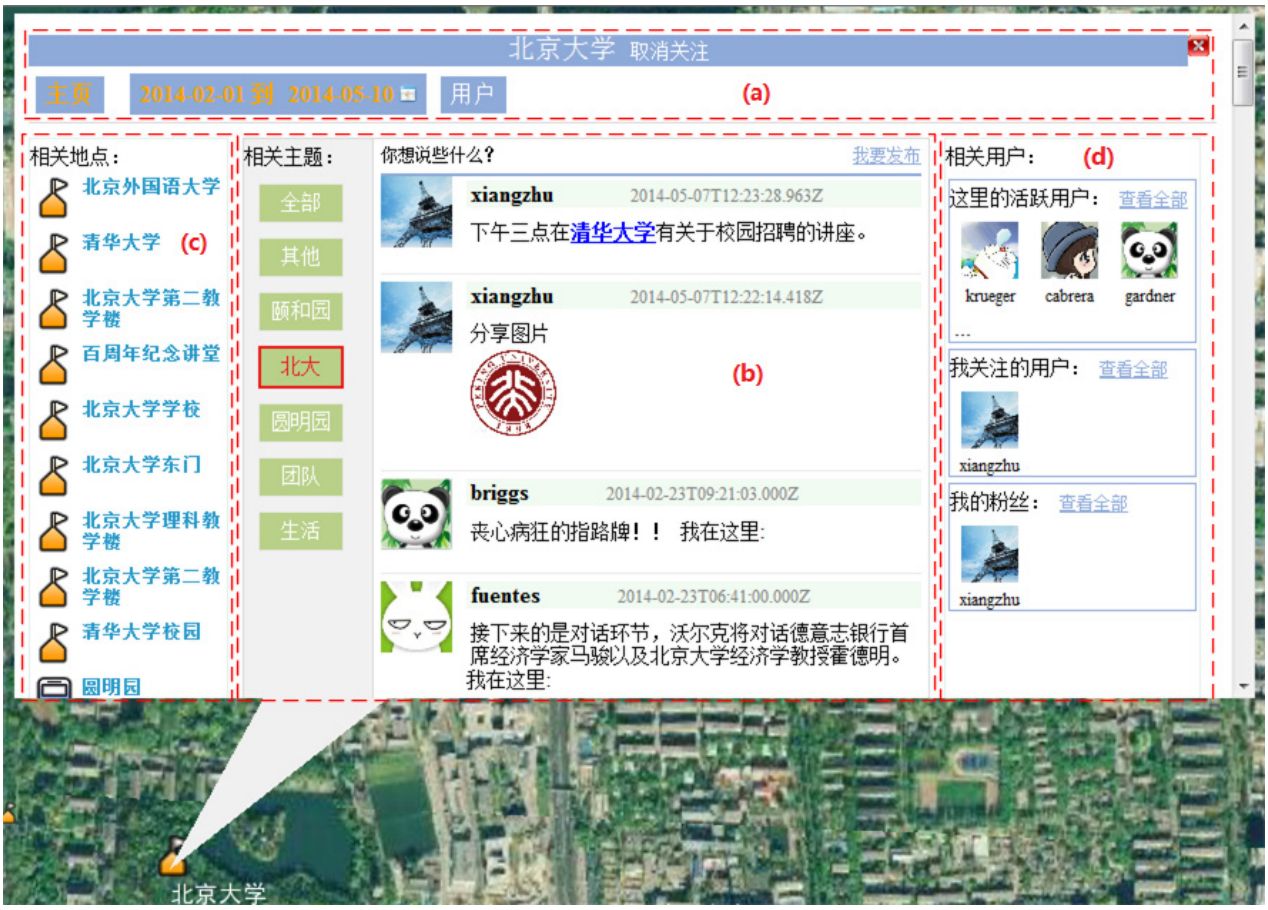

图 2　多线索的信息显示

当前线索. 特别地, 用户可以加入多个用户组, 每个用户组表示用户的一个社交圈, 本文的用户线索不仅提供单个用户, 也提供用户组的选择, 能够聚合呈现用户的某个社交圈的所有相关信息. 图 2 中 b,c,d 分别用于显示与当前所选线索相关的地点、消息和用户; b 部分同时提供当前地点上的主题选择, 点击 c 部分中的地点链接能够切换地点线索, 跳转到相应位置上查看信息.

3.3 基于位置的信息解聚显示

除基于兴趣点来发布和查看信息在某些情况下可能并不够理想, 用户不能最直观地看到地理位置上的相关信息, 本文设计了信息的解聚合显示方法, 将用户在地点上发布的信息直接利用“便笺”的形式显示在虚拟场景的相应位置上, 如图 3 所示.

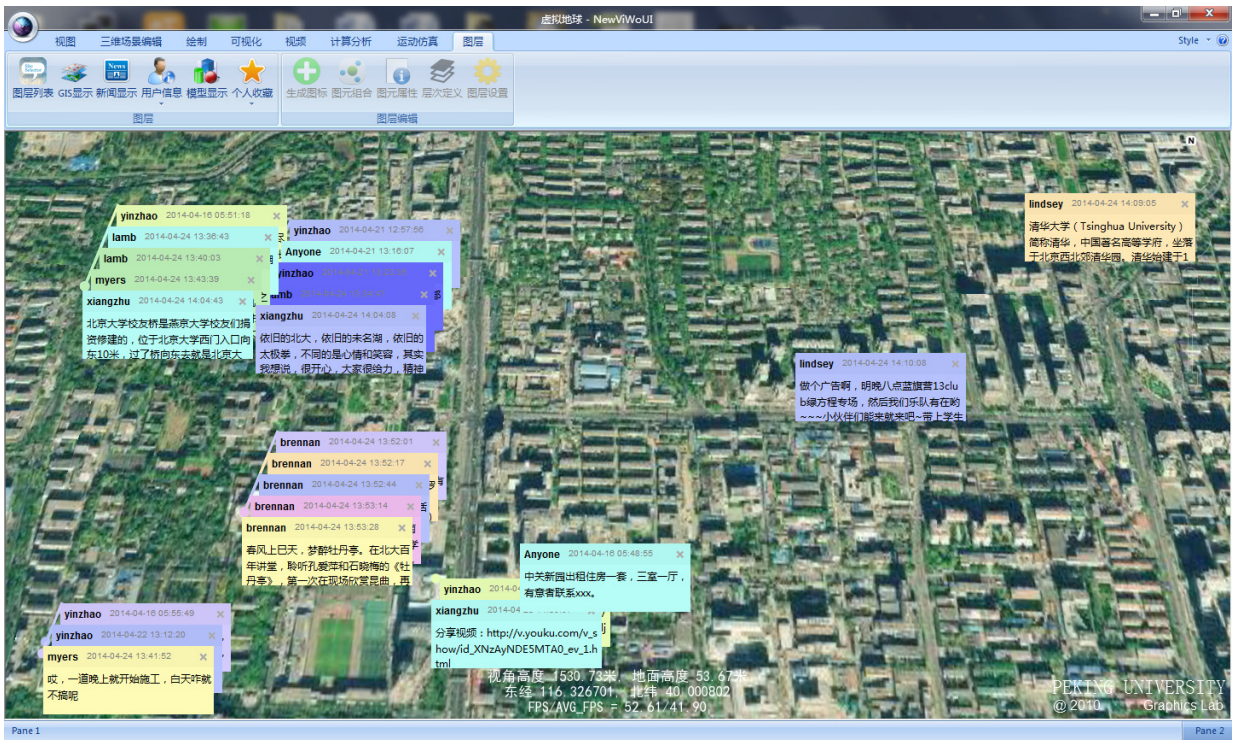

图 3 信息的解聚显示

每条消息的内容通过一个形似便笺的视觉元素来呈现, 便笺矩形结构的左上角顶点对应消息发布地点. 一个便笺的主要结构包括标题栏和主体 2 部分, 并通过具有一定色差的 2 种颜色区分. 为最大程度地利用便笺的空间, 标题栏上显示消息的发布者和发布时间, 主体部分则具体显示消息内容.

对于同一个地点(兴趣点)上的消息, 如果都按照原始的位置显示便笺, 它们会重合在一起不便于选择查看. 本文将同一个地点的若干个便笺以“堆叠书签卡”的方式显示, 除了最前端的第一个便笺屏幕位置与其原始地理位置对应, 其他便笺的屏幕位置依次偏移一定距离, 不再对应其原始地理位置, 通过一条线与原始地理位置连接, 并以点标志原始位置. 堆叠书签卡中的每个便笺透出其标题栏, 能够为用户提供概览信息, 也便于选中. 另外, 用户能够根据自己的需求和偏好自主拖动所有便笺, 拖动操作基于便笺结构的标题栏.

总之, 本文的可视化方法在虚拟场景中提供以地点为主线索的信息聚合呈现, 其中辅以用户、时间、主题线索, 用户能够通过自然的交互操作指定多种线索, 缩小相关信息的范围, 大大提高信息浏览效率. 同时提供信息的解聚合显示以便更直观地查看, 用户可以根据需求自由地切换信息的聚合和解聚呈现方式.

4 实验评估

本文设计了实验对主题提取算法进行了评估, 主要目的是评估主题词标注消息的精度及主题词的质量. 通过新浪微博提供的 API, 以“北京大学”为中心获取了其附近的 200 个兴趣点, 对每个兴趣

点，按时间范围抓取了该地点上 2014 年 1 月 1 日到 2014 年 3 月 31 日的微博数据，最终实验样本总量为 200 个兴趣点的共 128 044 条微博.

实验中，利用本文方法提取了每条微博的主题词. 目前本文并没有研究如何自动选取主题提取方法中的 3 个参数—LDA 的主题数目 K，每个地点的候选主题数目 K_c，判断主题语义强弱的概率阈值 p_t. 这些参数均通过多次实验选取其最优取值，即使得所得的主题词集合具有最强的语义可解释性. 在后续的应用中，可以视具体应用场景为它们赋以经验值. 在本文的实验中，设置 LDA 的主题数目 $K = 60$，每个地点的候选主题数目 $K_c = 6$，判断主题语义强弱的概率阈值 $p_t = 0.003$.

最后，采用人工评测的方法来评估得到的主题词是否符合微博的语义，从实验数据中随机抽取了 500 条微博作为人工评测的测试集. 经过对主题词抽取结果的预先分析发现，大多数微博的主题词被标注为“生活”，这类泛化的主题并不容易评测，因此最终的评测样本为随机抽取的 500 条主题词不为“生活”的微博.

实验 **1.** 评测者阅读测试集的每一条微博，并查看自动提取得到的主题词，判断该主题词是否能够较为准确地描述微博的语义并评分，主题词与微博语义相关较强评分为 1，主题词能够部分描述微博的语义评分为 0.5，主题词与微博语义无关评分为 0. 为帮助评测者理解，同时提供了每条微博所属的地点名.

实验 **2.** 本文将主题解释为主题词的方法较为简单，为进一步验证主题抽取算法的有效性，实验中为测试集的每条微博提供了其抽取得到的主题的权重最高的 3 个单词，若这 3 个单词中任何一个与微博语义相关较强则评分为 1，部分语义相关评分为 0.5，都不相关评分为 0.

共有 5 位评测者对上述两个实验的测试集进行了评测，他们都是微博的日常使用者. 表 1 展示了实验 1 和实验 2 的评测结果，本文以精度来量化描述主题词的质量，用百分比表示，本质上等于测试集所有微博评分的平均值.

综合对实验数据的分析、评测结果和评测者的反馈，本文得到了如下结论和一些启发：

表 **1**　实验的评测精度　　%

实验	评测 1	评测 2	评测 3	评测 4	评测 5	平均
1	45.0	49.2	42.6	39.2	37.2	42.64
2	52.6	56.0	48.2	51.0	47.6	51.08

1) 无监督的自动提取主题是一个较为困难的问题，例如 Bernstein 等[7]方法对于 Twitter 的主题精度为 40%. 在实验 1 的评测结果中，5 位评测者评估的主题词平均精度为 42.64%，此结果是可观的. 尽管不同评测者的评测标准各有不同，但精度基本在 40%左右，这也说明本文的主题词提取结果能够符合大多数用户的认知.

2) 实验 2 的评测结果中平均精度为 51.08%，显著高于实验 1，这进一步验证了本文方法的有效性，同时也体现出本文采用的 LDA 主题解释方法并不够理想，应当进一步研究如何将一个主题解释为最优的主题词.

3) 一般来说，用户发布的大多数微博可能只是简要地表达生活状态且内容较短，并没有明确的主题语义，对于在地点上发布的微博，更是不乏类似“我在北京大学，哈哈”这样的内容. 这样的微博并不容易标注主题词，这也是大多数微博会被标注主题词“生活”的原因.

4) 即使过滤了“生活”主题，测试集中仍然存在一些语义不明显的微博，对这些微博并不能很好地提取其主题，影响了最终的精度；而对于语义较为明确的微博，主题词的精度更加可观. 为语义不明确的消息赋予主题意义不大，需要研究一种消息预处理方法，首先过滤这类消息，仅对语义较明确的消息抽取主题.

5) 自动提取的主题词中有许多是地点名，如“北大”“清华”“中关村”等，这是由于用户在一个地点上发布的微博经常会提及该地点造成的，这类词具有一定的质量，也从侧面反映了以地点为单位抽取主题的合理性.

5　总　结

本文就如何对基于地理位置的社会信息进行有效的组织和合理的可视化进行了探讨，提出了一种多线索的信息组织方式，并针对其中的主题线索提出了一种自动提取消息主题的方法. 实验证明了这种方法的有效性. 本文基于多线索的组织方式，设计了一种在虚拟场景中显示社会信息的可视化方法，能够有效地提高用户的浏览效率，减轻用户交互负担，满足不同用户的需求. 尽管这种可视化方法是基于二维场景设计的，但能够方便地扩展至三维场景中，具有广阔的应用范围.

参考文献(References):

[1] Kwak H, Lee C, Park H, *et al*. What is Twitter, a social network or a news media?[C] //Proceedings of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 591-600

[2] Marcus A, Bernstein M S, Badar O, *et al*. Twitinfo: aggregating and visualizing microblogs for event exploration[C] //Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2011: 227-236

[3] Nichols J, Mahmud J, Drews C. Summarizing sporting events using twitter[C] //Proceedings of the ACM International Conference on Intelligent User Interfaces. New York: ACM Press, 2012: 189-198

[4] Zuo Y C, You F, Wang J M, *et al*. User modeling driven news filtering algorithm for microblog service in China[C] //Proceedings of the 11th IEEE/ACIS International Conference on Computer and Information Science. Los Alamitos: IEEE Computer Society Press, 2012: 393-399

[5] Chen J L, Nairn R, Nelson L, *et al*. Short and tweet: experiments on recommending content from information streams[C] //Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2010: 1185-1194

[6] Esparza S G, O'Mahony M P, Smyth B. Catstream: categorising tweets for user profiling and stream filtering[C] //Proceedings of the International Conference on Intelligent User Interfaces. New York: ACM Press, 2013: 25-36

[7] Bernstein M S, Suh B, Hong L C, *et al*. Eddi: interactive topic-based browsing of social status streams[C] //Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology. New York: ACM Press, 2010: 303-312

[8] Ren D H, Zhang X, Wang Z H, *et al*. WeiboEvents: a crowd sourcing weibo visual analytic system[C] //Proceedings of the IEEE Pacific Visualization Symposium. Los Alamitos: IEEE Computer Society Press, 2014: 330-334

[9] Tang J Y, Liu Z Y, Sun M S, *et al*. Portraying user life status from microblogging posts[J]. Tsinghua Science and Technology, 2013, 18(2): 182-195

[10] Gao T, Hullman J R, Adar E, *et al*. NewsViews: an automated pipeline for creating custom geovisualizations for news[C] //Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2014: 3005-3014

[11] MacEachren A M, Jaiswal A, Robinson A C, *et al*. Senseplace2: geotwitter analytics support for situational awareness[C] //Proceedings of IEEE Symposium on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2011: 181-190

[12] Yang Zhiqiang, Yin Zhao, Wang Heng. Providing resource recommendation based on interactive behavior and resource content[J]. Journal of Computer-Aided Design & Computer Graphics, 2014, 26(5): 747-754 (in Chinese)
(杨智强, 殷 钊, 王 衡. 结合用户交互行为和资源内容的资源推荐[J]. 计算机辅助设计与图形学学报, 2014, 26(5): 747-754)

[13] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3(Jan): 993-1022

虚拟场景中社会信息的组织和显示

作者： 殷钊， 王衡， 汪国平， Yin Zhao， Wang Heng， Wang Guoping
作者单位： 北京大学图形与交互技术实验室北京 100871; 北京市虚拟现实与可视化工程技术研究中心北京 100871
刊名： 计算机辅助设计与图形学学报 ISTIC EI PKU
英文刊名： Journal of Computer-Aided Design & Computer Graphics
年，卷(期)： 2015(10)