DOI:10.1145/2735624

**地铁图可以叙说故事，而不仅仅是指明方向。**

DAFNA SHAHAF，CARLOS GUESTRIN，
ERIC HORVITZ，JURE LESKOVEC

# 信息地图

“如果你没完全理解此次金融危机的整个过程，请举手”，David Leonhardt在2008年3月出版的《*纽约时报*》文章中写道。信用危机曾持续了七个月，世界各大媒体也对此进行了广泛和持续的报道。尽管存在这些深入报道，很多读者仍感觉他们并不了解这次危机的来龙去脉。

矛盾的是，铺天盖地的媒体报道可能是导致公众缺乏理解的原因，这一现象被称为信息过载。最近的技术进步能让我们以让人惊奇的速度产生数据。同时，网络的崛起还打破了传播的壁垒。然而，尽管数据洪流的速度越来越快，知识和注意力仍然是珍贵和稀缺的商品。作家、研究人员和分析家花了无数时间收集信息，整合出有意义的故事，同时检查和推导出各信息之间的关系。故事发展得越来越复杂，其中的细微处和关系在内容的不断修改和复用中很容易丢失，而吸引索引程序、眼球和广告点击的动机又加剧了这一情况。从大的数据集中自动抽取结构化信息的问题变得越来越普遍。

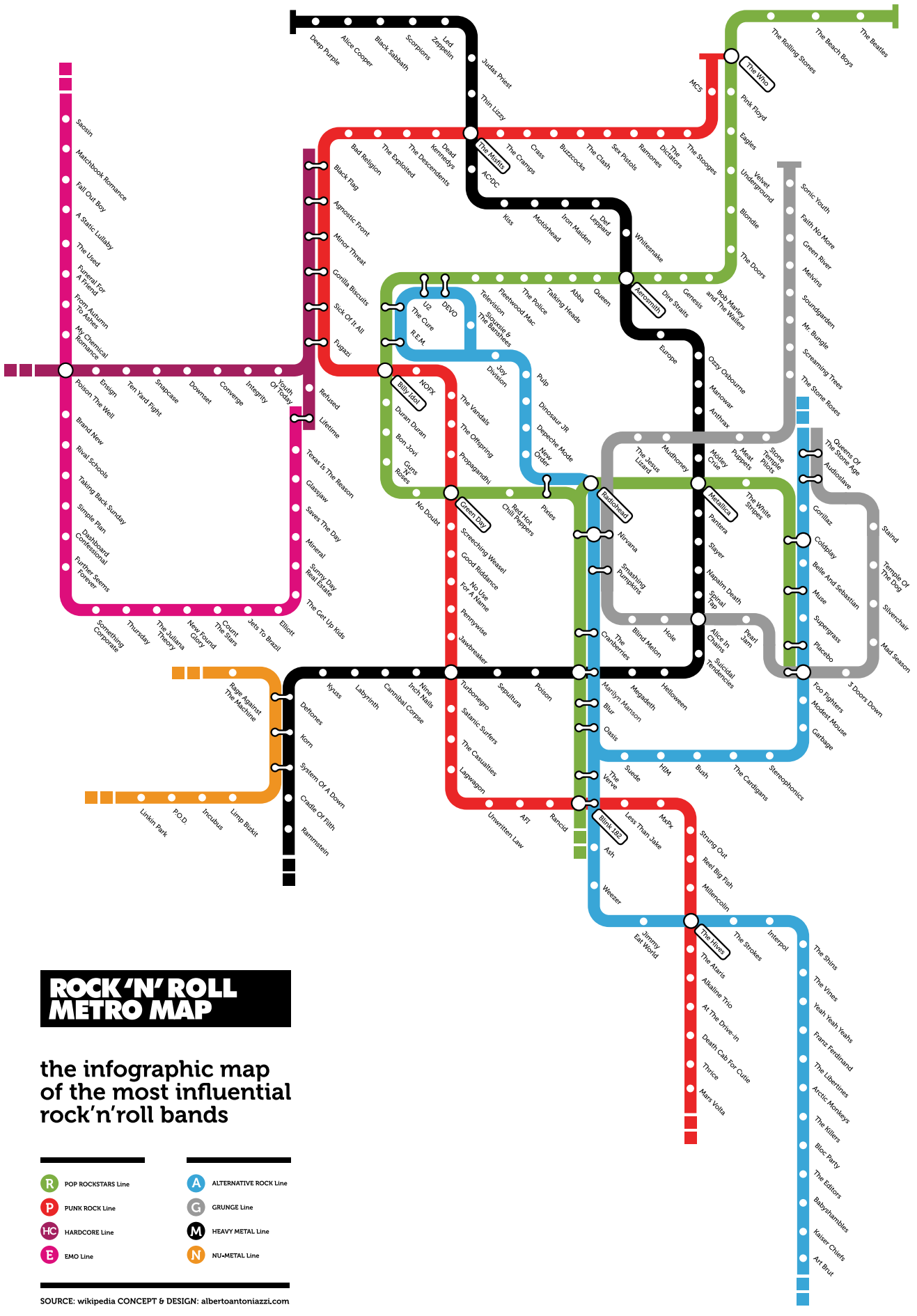
人们已经提出了几种汇总和可视化叙述故事的方法。[2,28,29]不过，大多数研究只能处理本质线性的简单故事。相比之下，复杂的故事展现了非线性结构；故事会变出很多分支、出现外传、走到尽头以及相互缠绕。为了探索这些故事，用户需要一张地图来引导他们穿过不熟悉的领域。

之前，我们介绍过一种创建结构化信息摘要的方法论，将其称为“地铁图”。该名称是隐喻性的；正如我们几百年来一直依赖地图来帮助我们理解周围环境一样，地铁图帮助我们理解信息的格局。在本文中，我们探讨了我们开发的，用于自动创建信息地铁图的多种方法。[25–27]

地铁图由互相交叉或重叠的线条集合组成。最重要的是，它们采用了一种可描述故事发展过程的方式清晰地展现了不同信息之间的关系。每个地铁站是一组文章，线条则追踪了连贯的叙事线索。不同的线条关注故事的不同方面；例如，图1中的地图是为查询“克里米亚（Crimea）”而自动生成的。该地图勾勒了2014年的克里米亚危机，其中的三条线条对应俄罗斯、乌克兰和西方的观点。各条线左侧的图例说明了每条线中的重要词汇。时间线标在图的底部。俄罗斯（绿）

## 重要见解

- **虽然人类的注意力和理解力可能会被数据洪流淹没，但是自动化的方法能够抽取结构化知识并提供描绘复杂信息格局的地图，帮助人们理解理念、联系和故事线。**
- **好地图的属性很难被形式化；重要的特性包括故事线的连贯性、对多样化和重要主题的覆盖以及信息之间的关系。**
- **这些原则可被用于从跨多个领域的大型数据集中汇总得出有意义的描述，这些领域包括新闻故事、研究论文、司法案例和文学作品。**




# ROCK 'N' ROLL METRO MAP

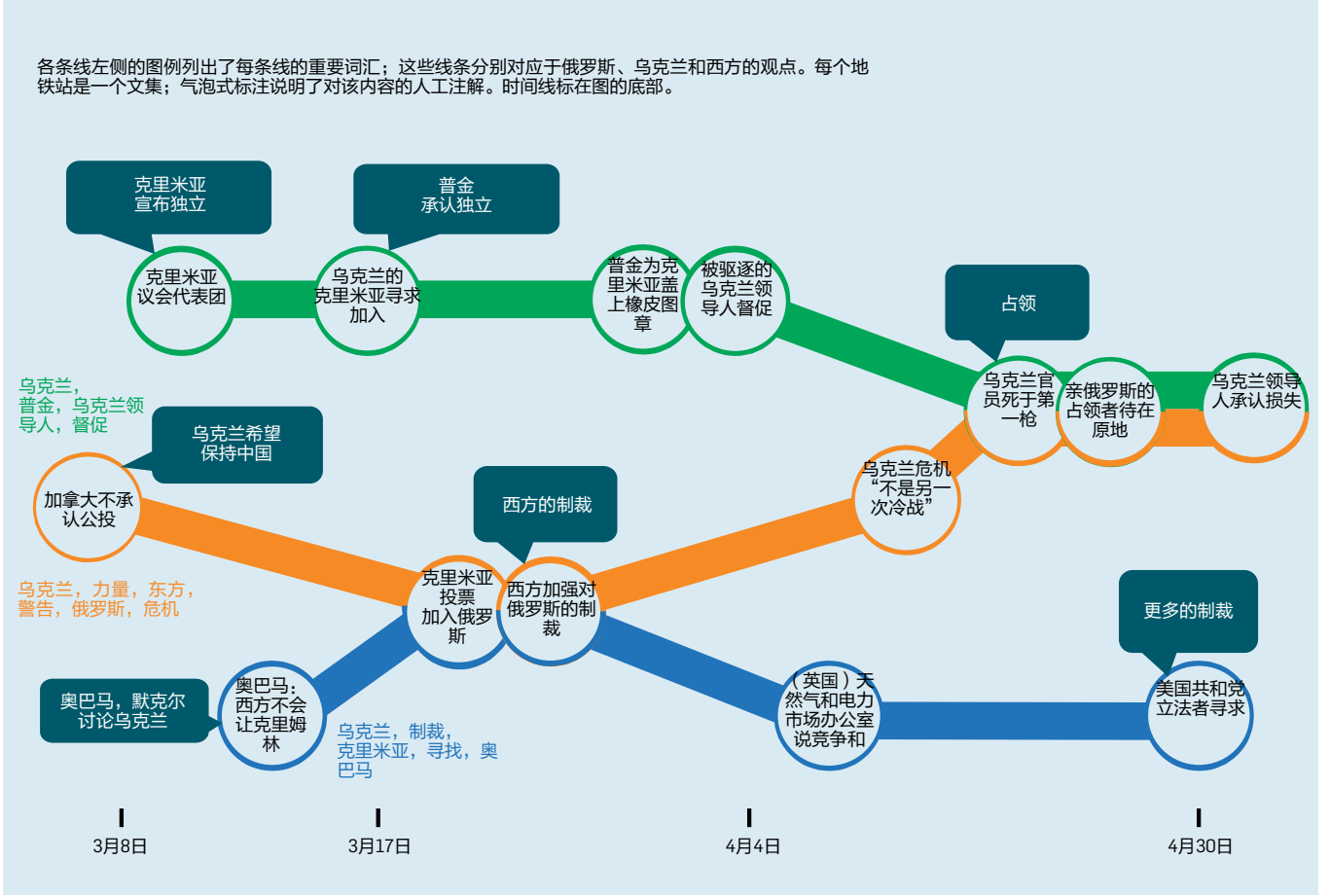
## the infographic map of the most influential rock'n'roll bands

- R POP ROCKSTARS Line
- P PUNK ROCK Line
- HC HARDCORE Line
- E EMO Line
- A ALTERNATIVE ROCK Line
- G GRUNGE Line
- M HEAVY METAL Line
- N NU-METAL Line

SOURCE: wikipedia CONCEPT & DESIGN: albertoantoniazzi.com

地铁图由ALBERTO ANTONIAZZI创建

**图 1.样例输出：2014年克里米亚危机的地铁图。**




线从三月开始，当时克里米亚议会投票加入俄罗斯，弗拉基米尔·普京承认克里米亚独立。乌克兰（橘）线从乌克兰前总理催促西方阻止俄罗斯入侵开始。然后，乌克兰线与西方（蓝）线交汇在一起，以讨论西方支持乌克兰的措施。最后，俄罗斯与乌克兰的线相交，当时亲俄罗斯的团体接管了乌克兰的警察站。

地图的表现形式能帮助用户获取和保留知识，而我们的这种表现形式就是受到这一有力经验证据的启发；例如，思维图和知识图已被证明可增强学生的回忆，[11,23]并提高他们的积极性和注意力。[15]我们还发现，利用地图进行可视化能让用户消化信息。我们还整合了多种能力，以便在方法论中融入用户互动，让用户引导地图的创建。

我们论证了地铁图能够帮助人们理解很多领域的信息，其中包括新闻故事、研究领域、司法案例、甚至是文学作品。地铁图能够帮助人们处理信息过载，确定了自动化信息抽取的一个研究方向，同时为汇总和展现由相关联概念组成的复杂集合指明了新的表现形式。

## 找到一张好地图

首先，我们形式化了好地图的特点，把构建好地图表示为一个优化问题。然后，我们为构建地图提供了高效的、可伸缩的方法，并带有理论保证。我们特意把特点描述得相当抽象。接下来，我们论证了如何改变这些抽象的概念，使它们适用于各种领域。

**目标函数。**在我们提出计算好地图的算法之前，我们必须构造一个目标函数。对于地图来说，这点特别重要，因为其中的目标并不清晰，是先验的。在后面的章节中，我们使用和形式化了几个（有时候冲突的）准则。在下一节，我们会提出一个构建地图的原理性方法，它优化了这些准则之间的取舍。

首先，回顾下我们的目标。给定一组文档，我们力求计算出一张汇总和组织这些文档的地铁图。地铁图由一组地铁线路组成，每一条线路包含一组有序的地铁站，每一个地铁站是文章的一个子集。每一条线路追踪了一个连贯的叙事线索，不同的线路关注故事的不同方面。线路之间的交汇说明了不同故事情节之间的互动方式；例如，在图1中，我们使用2014年3月到4月期间包含词语“克里米亚（Crimea）”的新闻报道计算出了一张地图。每个地铁站是一组文章集。该地图包含了三条故事线，追踪了俄罗斯、乌克兰和西方的观点。

*连贯性。*第一个要求是，每条地铁线讲述一个连贯的故事；沿某条线路的文章应该能让用户清晰地理解故事的发展。

考虑一下一条文集链，其中的每个文集均包含一组文章。为了便于表达，我们的重点放在单例上面，其中每个文集只包含一篇文章。为了定义连贯性，第一步自然是测量链上连续的每两篇文章的相似度。由于一次坏过渡可以摧毁整

个链条，我们通过测量链条中最弱环节的强度来得出链条的强度。

然而，这种简单的方法可能会产生相当差的链条。比如，考虑一下链A和链B。它们有相同的端点，但链A的连贯性明显要低一点。注意，离开具体场景观察时，链A之间的所有转换都合理；前两篇文章与债务违约有关，第二篇文章和第三篇文章与共和党有关，等等。虽然存在这些局部连接，但整体的效果是不一致的。

现在，我们来仔细看下这两条链。图2说明了两条链中出现的词语；例如，词语“希腊”一直在链B中出现。发现链A的关联流也相当容易。词语会在链条的一小段中出现；有些词语会出现一下，然后消失，然后再次出现。相比之下，链B中的小段更长，转换更平滑。这一观察推动我们做出了连贯性的定义。

我们把该问题变成了一个线性规划优化问题，其中的目标是选择一小组词语，然后只根据这些词语来为链条评分。为了确保每次转换的强度，链条的分数（给定活跃词语的集合）是最弱环节的分数；详细信息参见Shahaf 和Guestrin[24]的论文。

单一环节的分数可能依赖于领域。在Shahaf等人[26]的论文中，我们说明了在只给出文章内容时，如何计算分数。在Shahaf等人[25]的论文中，我们说明了如何利用文章之间的关联。

*覆盖面*。连贯性是好地图的关键因素，但它是否足够呢？为了得到答案，我们找出了查询“比尔•克林顿（Bill Clinton）”时，连贯性最大的线路。结果令人沮丧。虽然这些线条事实上是连贯的，但是它们不重要。很多线路围绕着狭窄的话题发展（比如克林顿访问贝尔法斯特）。不仅如此，因为没有多样性的概念，多条线条包含了冗余信息。该例子说明，选择最连贯的线条并不能保证一张好地图。与此相反，关键的挑战是平衡连贯性和覆盖面；除了连贯之外，线路还必须覆盖对用户重要的广泛主题。

我们定义了地图能够覆盖的元素集合。这些元素可能依赖于领域；在新闻报道的情况中，我们选择了词语（比如“奥巴马”和“中国”），[26]高覆盖面的地图会讨论很多重要的词语。在科学语料库场景中，我们选择了论文[25]，所以高覆盖面的地图会触及语料库的很大一部分。

我们计算了覆盖面函数，测量了每篇文章对每个元素的覆盖情况。我们把它扩展到了集合函数，测量了文章集合对每个元素的覆盖情况。为了鼓励多样化，该函数使用了次模函数；如果该地图已经良好覆盖了某个元素，那么加入良好覆盖该元素的另一篇文章几乎不会提高覆盖面。相比之下，由于无法提高覆盖面，我们更愿意选择覆盖新主题的文章。

然后，我们为每个元素引入了权重，用于说明该元素的重要性。权重让地图偏向于覆盖重要的元素，同时也为个性化提供了一个自然的机制。在Shahaf 等人[26]的论文中，我们讨论了如何从用户反馈中学习权重，进而得出个性化的覆盖面概念。

*连通性*。最后，地图不仅仅是一组线路，它还有结构信息。因此，我们最后的属性是连通性。地图应该传递出故事的底层结构以及故事的不同方面如何相互影响。

从直觉上来说，不同的故事有着不同的结构。有些故事大致是线性的，而其他的故事则复杂得多。为了捕获故事的结构，我们计算了覆盖所有地铁站点的最小线路数。该目标偏爱尽可能地选择长故事线；线性故事变成了线性地图，复杂的故事保持了它们互相交织的线索。

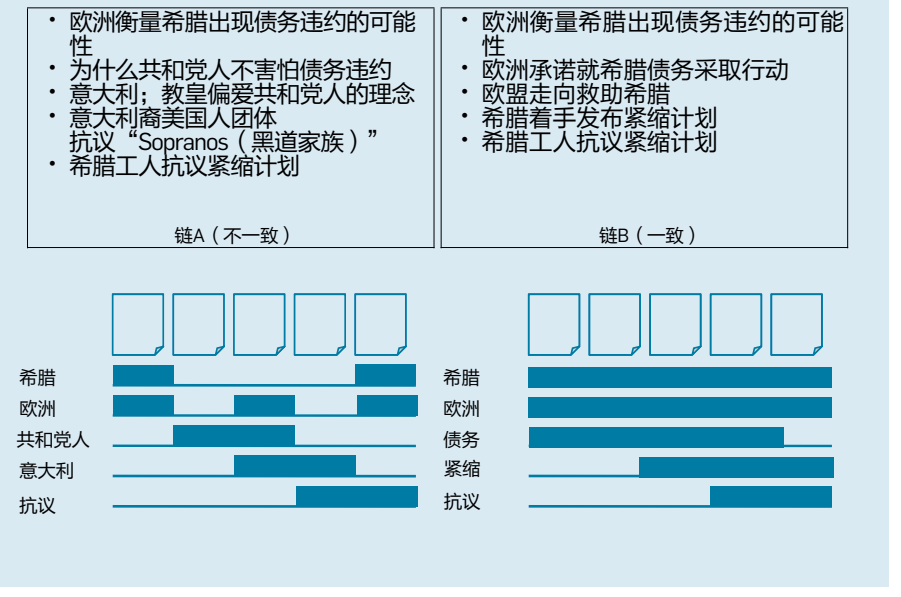
*把它合在一起*。现在，我们来用公式表示在给定一组文章时如何找出好地铁图这一问题。我们需要在之前讨论的各种属性之间做出取舍：“文集质量”、“线路连贯性”、“地图结构”以及“符合预算的覆盖面”；例如，最大化覆盖面会导致不连通的地图，因为没有理由在多于一条线路中重用某个文集。最大化连贯性往往会导致重复的，范围狭窄的链。因此，最好把连贯性当成约束；链的连贯性或是达到了可让其在地图中出现的程度，或是没达到这种程度。另一方面，覆盖面和结构均应进行优化。我们将地图的目标定义如下：

*问题1（地铁图:非正式）*

地图必须满足

| | |
|---|---|
| 高覆盖面 | (*o1*) |
| 高结构质量 | (*o2*) |
| 前提条件是达到 | |
| 最低的线路连贯性水平 | (*c1*) |
| 最低的文集质量 | (*c2*) |
| 最大的地图尺寸 | (*c3*) |

**图 2.链A（左）和链B（右）中的词语模式；柱形对应词语是否在上文列出的文章中出现。**



该算法及其优化的形式化描述参见Shahaf等人[27]的论文。

## 算法

现在，我们简短地回顾下该算法背后的主要理念。首先，我们从计算查询时的文章集合开始。然后，我们把文章分别放在不同的时间窗内，使用处理词共现图的社区探测算法计算每个时间窗中的好文集（问题1中的约束$c2$）。[27]把这些文集作为地铁站。

一旦我们有了文集之后，我们便能够继续下去，计算连贯性线路（约束$c1$）。理想情况下，我们能够计算出所有可能的备选线路，但是这往往不可行。相反，我们提出了分治方法，用较短的线路构建长线路。该方法能让我们把很多条线路紧凑地编码在一张图内；图中的节点对应短的连贯线路，图中的边说明可以级联并保持连贯性的线路。因此，图中的路径对应连贯的线路。

为所有连贯的线路编码后，我们确定了故事的底层结构，优化了连通性目标（$o2$），即尽量选择更长的故事线。该目标确实很难优化，但它是次模的，可以在保证的范围内有效逼近。

**图3.2013年5月查询“Boston”时得出的地铁图**

两条线讨论了波士顿马拉松爆炸的后果，其中一条线关注嫌疑人，另一条线关注社区事件；其他的两条线则与波士顿的大联盟运动（曲棍球和棒球）有关。

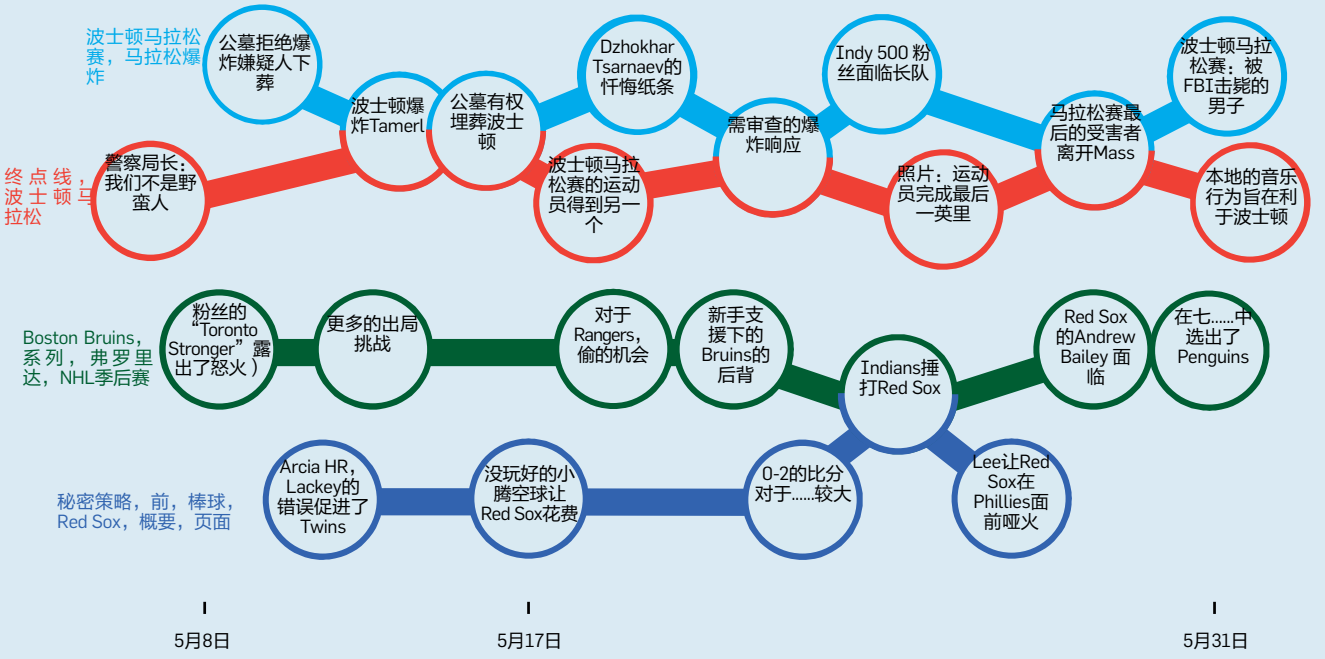
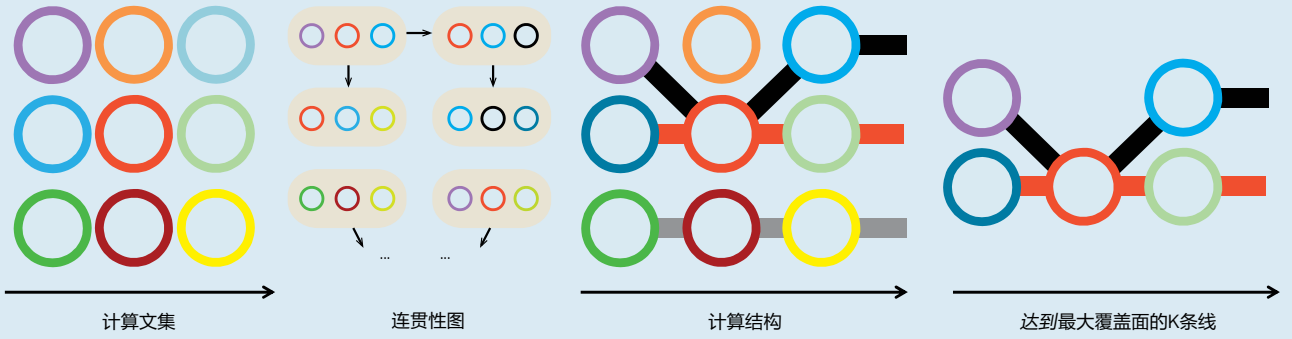



**图4.算法概述。我们计算出了文集，在图中对连贯的线条进行了编码，并使用图来计算故事的结构。然后，我们从最大化覆盖面的结构中选取了$K$条线。**



故事可能相当复杂，但用户的关注范围是有限的。为了让地图一直易于管理，我们最后的步骤是限制地图的尺寸。我们会选择最多$K$条符合结构的线路，然后最大化覆盖面（$c3$和$o1$）。我们再次依赖次模优化在理论保证的范围内优化地图。图1（克里米亚）和图3（波士顿）说明了该算法的样例输出。

*复杂度和运行时间。*给定文章的查询集$D$，我们首先运行一个线性时间的算法，把$D$编成一系列词共现图。更重要的是，图的大小不依赖于$D$，但依赖于我们的词汇$W$的规模。我们对$D$的规模的依赖是线性的，而且我们的算法具有良好的伸缩性（见图4）；理论保证参见Shahaf 等人的两篇论文[26,27]。我们主要的瓶颈卡在覆盖面的步骤，它是$|W|$的高阶多项式。采用并行实现和延迟求值后，可达到相同的逼近保证，而且速度常常能获得巨大的提升。

在实践中，对于含数万文档的查询集，我们的系统所花时间往往不到一分钟。注意，虽然我们的系统原则上支持更大的查询集，但我们心中的地图用例极少需要这么做。我们设想，非常宽泛的查询（“美国”）较范围更窄的查询【“医疗改革（health care reform）”】更不常见。

*参数。*为了获取好的地铁图，需要对几个参数进行调优。具体来说，问题1中的每个约束均有其关联的参数，需要在训练查询方面进行手动调整。另一个重要的参数是$m$，即用户的“历史窗口”，或者说该线路中用户能够记住的，以前的文章数量。$m$越高，链条的连贯性也越高，但是计算的代价也更高。在实践中，我们选择了在计算上能够承受的最高的$m$值。

## 应用

在之前的章节中，我们讨论了新闻领域内的地铁图。不过，我们可以轻松地把地铁图应用到其他领域。主要的原则——连贯性、覆盖面和连通性——是相同的，但是人们可以利用领域知识来改进目标。在接下来的章节中，我们讨论了四种应用：新闻、科学、法律文书和书籍。

**平衡；线用题是盖外盖泛主战覆之覆广挑和贯须的性连必要的关键贯了还重条除连户。**

**新闻。**在向公众传播社会、文化和政治问题的过程中，新闻媒体发挥了关键作用。理解新闻能让公众做出一生中的关键决定（比如选择居住地或政治倾向）。在没有洞悉全局的情况下行动，后果可能适得其反。然而，由于每天出版的内容量越来越大，读者可能很容易在数据洪流中错失全貌。

*方法。*我们使用了我们的算法来计算新闻事件的地图，其汇总了从互联网新闻数据源中获取的，覆盖数十万博文的新闻数据集；该系统的演示参见http://metromaps.stanford.edu/。

*评估。*对地铁图进行量化评估相当困难。在量化评估方面，缺乏已经成型的黄金标准，同时又很难定义真实值。因为地图的目标是帮助人们浏览信息，所以我们开展了一次用户研究，以更好的理解该方法论的价值。该研究于2011年在宾夕法尼亚州匹兹堡市卡内基梅隆大学举行，旨在检验我们生成的地图是否能帮助人类寻找与复杂话题有关的信息。

为了说明用户对话题的理解深度，我们要求用户向他人解释该话题。我们招募了15名本科生，要求他们写了两段文章，一段总结了海地地震的情况，另一端总结了希腊的债务危机。我们随机地给他们分配地铁图或谷歌新闻的结果页。我们利用由*《纽约时报》*的18，000多篇文章组成的语料库计算出了地图。我们雇佣了Mechanical Turk（http://www.mturk.com/）上的众包工作者评估这些段落。在每个回合中，我们会给众包工作者展示两个段落（地图用户vs. 谷歌新闻用户），然后要求他们评估哪个段落提供的故事信息更完整，更连贯。删除了垃圾信息后，在希腊方面，我们得到了294份评估结果；在海地方面，我们得到了290份评估结果；希腊方面，72%的对比结果觉得地图段落更好；海地方面，只有59%的对比结果觉得地图段落更好。在检查海地方面的段落后，我们发现大多数地图用户只关注了（与派发救援物品有关的）主故事线。根据该研究结

果，我们推断，如果故事中没有单一的，可支配全局的故事线，地图的作用更大。

**科学。**由于科学出版物的数量飙升，即使对于最热心的读者，他们想站在不断发展的文献前沿也有困难。我们的动力源于我们希望创建宝贵的，能帮助人们（比如跨越传统学科边界的研究生或专家）进入新领域的文献探索工具这一理念。

*方法。*我们把我们的技术扩展到了科学领域，旨在检验地图是否能够帮助研究人员理解某领域的最新进展。我们利用引用图的优点稍微修正了这一目标（参见Shahaf等人的论文[25]）。我们的数据集包含了源于ACM会议和杂志的35,000多篇文章。

图5勾勒出了我们为查询“reinforcement learning（强化学习）”而计算出的地图的一部分。图中展示了多条研究线路，其中包括马尔可夫决策过程、机器人和控制、边界及其分析、探索-开发均衡（exploration-exploitation tradeoff）以及多主体协作（multiagent cooperation）。注意，图中的线条并不相交。在科学领域内，相交并不容易；理论线和应用线可能高度相关，但是没有哪一篇文章同时属于两者。因此，我们修改了我们的目标，允许更弱的连通性类型，其中的线条可以通过引用相互影响；例如，该地图说明了马尔可夫决策过程如何影响多主体和机器人线以及探索-开发线如何与分析线相互影响。这些关系通过灰色的虚线路径表示，附近标有相关的引文。

*评估。*为了检验我们的地图，我们从卡内基梅隆大学招募了30名研究生，要求他们快速地浏览强化学习的文献。他们之前并没有研究过这一领域。具体来说，我们要求他们确定最多五个研究方向，借此来更新1996年以来的论文综述。综述中应该包含这五个研究方向，并在每个方向上列出一些相关论文。我们记录了他们的浏览历史，同时也记录了他们每分钟的进展快照。我们把他们的时间限制在40分钟内，以模拟首次快速浏览论文的情况。所有的参与者均使用了Google Scholar（谷歌学术）[a]。Google Scholar是对学术文献进行索引的搜索引擎。另外，我们给了其中15个人一张地铁图。我们允许他们查询Google Scholar的整个出版物集合，这不仅让任务更贴近现实，还增加了地图处理任务的难度。

**地图被设计用于展现多个信息之间的联系。**

一名专家会给所有参与者的输出打分。我们要求他们找到好论文，同时确定重要的研究领域。因此，我们测量了精度（检索出的相关文章的比例）以及副主题的召回率（检索出的相关研究领域的比例）。在每一个参数中，地图用户的表现均优于只使用谷歌的用户，他们的平均分数为84.5%，平均发现了1.62 篇开创性论文。谷歌用户的分数达到了 74.2%，平均找到了1.2篇开创性论文。地图用户的平均召回率得分为73.1%。相比之下，谷歌用户只有46.4%。

在研究中，我们对快照进行了进一步分析，发现了地图用途的轶事证据。平均而言，谷歌用户访问的页面更多，列出的论文也更多。然而，当我们观察平均的比值时，对于谷歌用户，他们访问的每4.5个页面中只有一个被加入了他们的列表；对于地图用户，每3.8个页面中有一个被加入列表。也就是说，地图用户看起来更为聚焦；他们访问的页面可能更少，但是他们发现这些页面足以令人满意。另外，几个地图用户一开始会编写一个研究方向的短名单，然后在整个过程中逐步地向每个方向添加论文。相比之下，谷歌用户并没有表现出这种“大局观”的行为。

**法律文书。**法律基于不断发展的理念构建，与关键的先例之间存在关联。法律学者和律师会例行地研究法律文集，处理雪崩般的信息。虽然信息方面存在相关的挑战，但是法律文书和评审过程在很大程度上仍未受到技术的影响。我们试图探索地铁图在帮助律师辩论案件方面的价值。我们设想了一个系统，可以帮助他们找到相关的案件，理解法律的发展进程（以及修

a http://scholar.google.com

订法律的原因），然后准备相应的案件策略。

*方法。*我们的数据由Ravel Law提供的美国联邦最高法院判决组成。[b]与新闻报道和科学论文不同，美国联邦最高法院判决可能相当冗长，长度可达数百页。对于法律学者而言，前面章节中的简单文本处理方法可能无法分清良莠。

为了处理这一挑战，我们把重点放在了锚文本，即引文周围的文字上。确定高引用的段落后，我们便可以把重点放在每个案件中的重要部分上面；例如，*Women's Community Health v. Cohen*援引了*Roe v. Wade*，文字如下："最高法院认为，宪法中的"隐私权 ... ... ...的范围足以覆盖妇女是否终止妊娠的决定"（Id. 410 U.S. at 153 93 S.Ct. at 727）。我们使用锚文本来计算我们的输入文档集合，然后再使用我们的地图算法。

*评估。*为了进行真实性检查，我们计算了查询"commerce clause（商业条款）"的地图。它出现在美国宪法第1条第8款，文中规定国会有权"管理与外国的、州与州间的，以及对印第安部落的贸易。"这一条款相当重要，多家法院曾详细讨论过该条款。

图6（左）展示了我们计算出的地图。我们给Ravel Law的律师看了这张地图，请他们进行解释。他们浏览了该地图，人工标注了每一条线路。为了进行真实性检查，接着我们计算了让每条线路在我们的算法中保持连贯的词语。图6（右）展示了对比情况。连贯的词语与律师的标注对应得很好；例如，紫色的线路处理了国会是否会废除第十一修正案授予各州的豁免权这一问题。律师们把该线路标记为"第十一修正案，州主权"，它的连贯词语为"豁免权"、"主权"、"修正案"和"第十一"。

接着，我们要求律师们解释每一条线路。举例来说，再考虑下图6中紫色的线路。该线路从*Ford Motor v. Dept. of Treasury*开始。最高法院认为，第十一修正案没有授予联邦法院在未得到各州同意时受理私人诉讼该州的案件。

在下一站（名为*Parden v. Terminal Railway*的案件）中，最高法院讨论了拥有铁路的州能否在其雇员提起的联邦法院诉讼中使用主权豁免抗辩。在*Employees v. Dept. of Public Health*中，最高法院记录到，即使最高法院推翻自己的判决，国会仍能授权州法院处理可施行联邦法定权利的诉讼，据此可以废除习惯法中各州的主权豁免。接下来，在*Quern v. Jordan*中，最高法院认为，在对各州的诉讼方面，国会无任何意愿把各州包含在术语"人"的范围之内。最后，在最后一站（名为*Welch v. Texas* Highways的案件）中，最高法院否决了与各州参与联邦支出项目有关的，*Parden*案件的一个判决。

同样地，律师们能够解释所有其他的线路，表示出他们相信这些地图有利于法律社区。

b http://www.ravellaw.com

**图 5.查询"强化学习"时计算出的地图的一部分**

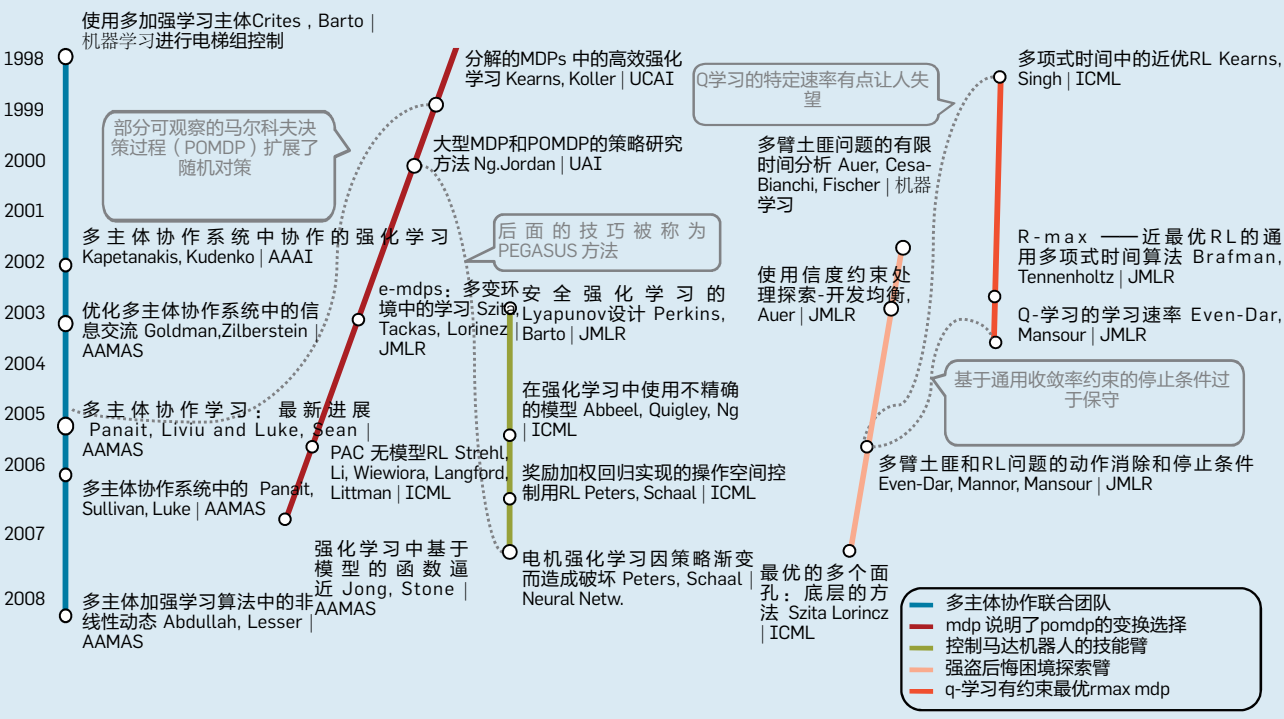
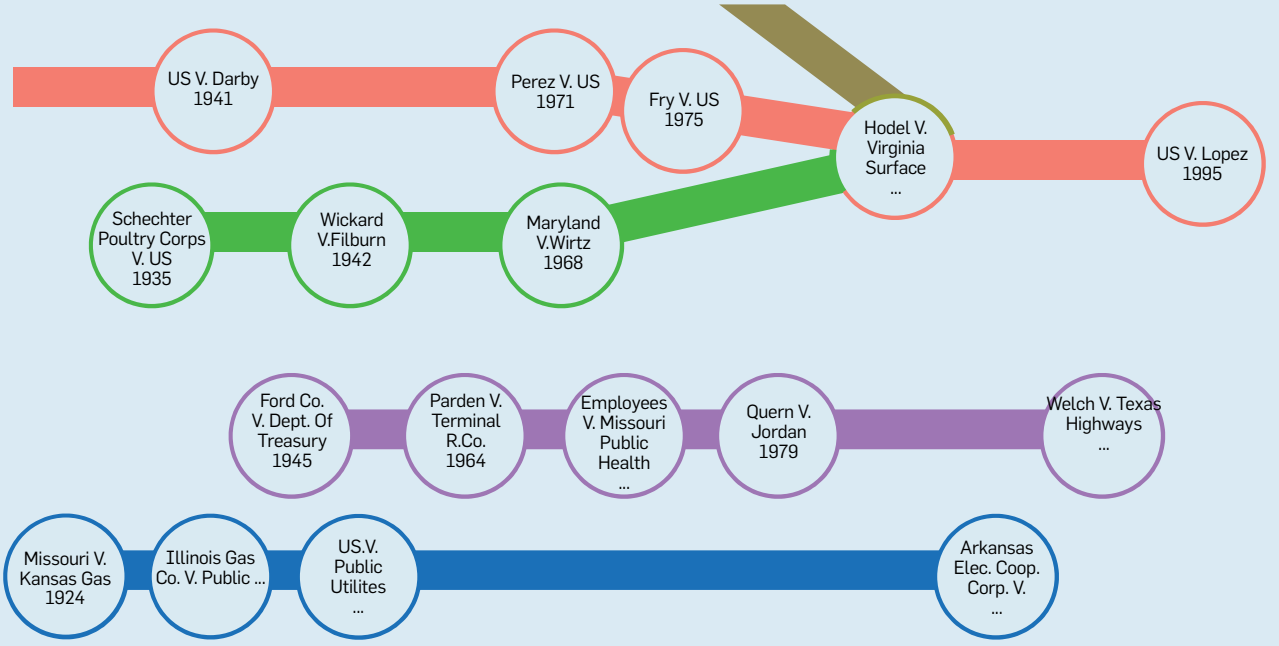


**图 6.查询”贸易条款（Commerce Clause）”时得到的法律文书方面的详细地图**

线条关注了美国国会禁止贸易的权力，第十一修正案，对能源批发销售的管制等。右侧：律师们对每条线路的标注，与我们的连贯性算法选择的词语进行了对比。对应于连贯性词语的人工标注。




| 律师们的标记 | 连贯性词语 |
|---|---|
| 禁止贸易的权力 | 州之间，贸易，影响，管理 |
| 国会管理的权力 | 国会，利益，管理，渠道 |
| 第十一修正案，州主权 | 豁免权，主权，修正案，第十一 |
| “仅仅” 对比 “实质性” 影响 | 影响，实质性，管理 |
| 管理能源的批发销售 | 批发，电力，转售，蒸汽，公共事业 |

**书籍。**在很多领域中，叙述相当重要。这些领域包括文学批评、政治科学和语言学。尽管如此，我们对它们的结构基本还是一无所知。因此，我们决定应用地铁图来阐明复杂书籍的结构。我们的第一个测试用例是《*指环王*》，这是一部史诗级的奇幻小说，字数超过480,000词，内部分了六卷。它里面包含了相当长的角色列表，即使是最专注的读者可能也很难追踪所有的角色。

方法。因为我们的地图在文章集合上运行，所以我们把该书分成了多个长度为三页的片段，把每个片段当成一篇文章。把地图算法应用到该片段集合上后，我们碰到了与连贯性概念有关的，有趣的问题。它最初出现在新闻场景中。记者往往会提醒他们的读者之前发生的事件，因此可以通过重复来推断连贯性。相比之下，书籍的作者并不会重新叙述在几页之前发生的事件。与此相反，他们依赖全神贯注的读者记忆。因此，我们依赖其他的提示来获取连贯性。

在注意到角色的观点往往是连贯的叙述之后，我们决定把重点放在命名实体上。我们确定了每页中出现的角色，然后寻找由一起出现的角色构成的故事线。

*评估*。图7展示了《*指环王*》地图的一部分，揭示了重要的结构信息：故事从夏尔（Shire）开始（最左边的文集），Gandalf 在那儿碰到了霍比特人（hobbits）。在与该文集相关联的书页中，Gandalf 建议Frodo 带着戒指离开夏尔。在第二个文集中，弗罗多离开了，陪着他的有Sam、Merry和Pippin。他们带上了Strider（后来发现是Aragorn）作为向导和保护者。

在接下来的文集中，Aragorn 把霍比特人带到了Rivendell（瑞文戴尔），在那儿召开了埃尔隆德会议。会议决定必须摧毁戒指，

并组建了“护戒使者（Fellowship of the Ring）”团队，其中包括Sam、Merry、Pippin、Aragorn、Gandalf、Gimli、Legolas和 Boromir。

会议的文集分出了三条线路（如图7所示），对应于半兽人（orcs）绑架Merry 和Pippin（绿线）时护戒使者团队分成几个小队。Aragorn、Gimli和Legolas追踪半兽人（紫线），Frodo 和Sam继续自己的任务，抓住了Gollum（蓝线），然后朝Mordor前进，即Sauron 和他的仆人Saruman 控制的区域（黄线）。

虽然这个例子比较初级，但它说明了地图整理复杂情节线的潜力。

## 使用地图

我们已经讨论了创建地图的过程，现在我们转向用户，探索地图的潜在用途。我们依赖传统的信息检索框架，通过某个用户的信息需求来刻画这个用户。用户会拟定他们的信息需求，然后向系统提交查询。如果他们对结果不满意，他们通过修改查询与系统交互，直到他们得到满意的结果。在本节中，我们讨论信息需求和交互的场景。

**信息需求。**我们并不打算用地图来代替搜索引擎；搜索引擎的很多查询非常聚焦，用一个简单的查询往往就能满足对应的信息需求。相比之下，地图被设计用于展现多个信息之间的联系。根据Broder的分类法，[7]信息类查询时使用地图的作用最大，导航类查询和交易类查询时使用地图几乎没用。

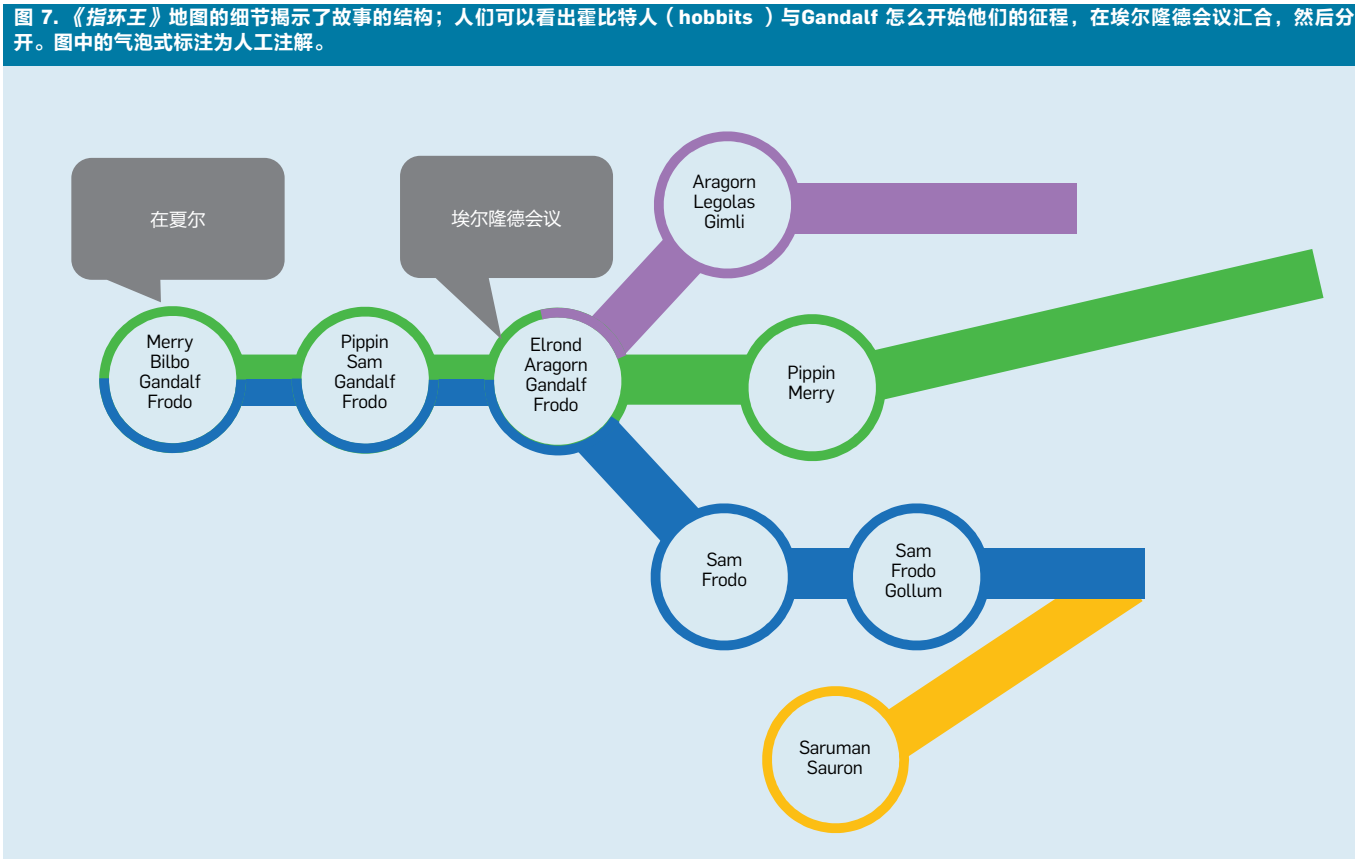
我们想刻画地图用户的信息需求。信息类查询的驱动力源于用户想学东西。为了刻画不同的学习类型，我们使用了Bloom的分类法[3]。该分类法确定了刻画学习过程的六个认知类别，其范围从回忆事实一直延伸到做出判断。

我们区分了地图用户的两类信息需求。“学习”对应Bloom分类法中较低的三个层级；该用户的目标是获取知识。学习类别中的用户可能对概览某个具体的主题感兴趣。对该用户而言，该主题可能是新的主题，或是该用户希望密切关注的熟悉主题。或者，用户可能希望从某个起点进行探索和浏览。导航是前景广阔的地图应用，现在有很多新闻网站包含了“相关文章”的功能。地图可以增强这一功能，让用户能在更宽广的背景下阅读文章。注意，学习类别不包括查找事实或回答问题。如Shahaf 等人[26]的论文所述，对于这种类型的查询，地图的作用要小一点。

“调查”类别对应于Bloom分类法中的较高层级。在这个类别中，用户的目的是得出结果。在该类别中，用户的目的是把已有的数据变成有用的模式，同时找出现有知识的缺口。他们分析和整合不同的信息，寻找可行的，可引发新洞见的概括知识。具体来说，此类用户的兴趣可能是对比和比较多张地图。

**交互。**交互是地铁图成功的关键。用户往往知道自己要找的确切信息，但是要他们把想法归结为几个关键词却不容易。因此，地图应该支持交互。很多交互模型可以被自然地整合到地铁图中。我们依赖用户反馈来学习用户偏爱的因素，然后相应地调整我们的地图。在接下来的章节中，我们会讨论我们已经实现的两种交互机制：“缩放”和“词语反馈”。

**图 7.《指环王》地图的细节揭示了故事的结构；人们可以看出霍比特人（hobbits）与Gandalf 怎么开始他们的征程，在埃尔隆德会议汇合，然后分开。图中的气泡式标注为人工注解。**



*缩放*。有些用户喜欢快速地获取某个主题的高层级概述，其他的用户则希望深入细节。因此，我们希望我们的地图是可缩放的。由于地图的表达能力相当强，缩放交互的解释可采取多种方式。我们已经实现了三种缩放解释方式：缩放可以影响时间分辨率；文集的分辨率会导致文集拆分和合并；用户可以聚焦于某条特定的地铁线上。

*词语反馈*。当用户与地图交互时，理想情况下，地图应该支持如“告诉我欧盟如何应对危机”或“我对Red Sox（红袜队）不感兴趣”这样的反馈形式。即使标注整个地图，或是单独的每一篇文章，得到的信息也不足以支持这种交互模型。而且，无法知道用户是否喜欢地图上的某个东西。

我们提出了“基于特征的反馈”，而不是提供一种自然的方式来支持前面讨论的查询；用户可以增加词语“欧盟（E.U.）”的重要性，降低“棒球（baseball）”的重要性，借此来达到期望的效果。我们使用了判别式半监督学习方法，其中融入了特征和类别之间的训练密切关系（training affinities）。[9]使用该方法后，我们定义了一个个性化的、对会话敏感的覆盖面概念，用于说明用户反馈。

## 相关研究

据我们所知，自动化构建地铁图尚属一个新领域。然而，在无数的相关研究方向中，研究人员已经开展了大量研究工作，范围从主题检测和跟踪延伸到摘要和时态文本挖掘。

与之前的研究相比，我们的研究存在以下几点重要区别：首先，我们的系统拥有结构化的输出，所以它不仅找出了信息块，还清楚地说明了它们之间的联系。之前的工作大都局限于列表输出的模式。在摘要任务中，[2,20,22]目标往往是通过抽取句子列表来得到文本语料库的摘要。其他的方法[18,31,30]发现了新的事件，但并未设法把它们串在一起。

在信息检索领域，研究人员投入了无数的精力来超越列表，以提供内容更丰富的视图，其中包括不同的故事线概念。[1,2,28,29]在从主题演化到新闻分析的多个相关问题中，[10,14,17,19]图的表现方式已经相当普遍。然而，在所有这些方法中，均没有使用图中的路径来表示连贯的故事线这一概念。与此相反，之所以使用图中的边，原因可能是它们超过了某个阈值，或是属于某棵生成树。

**给定某个查询后，我们的算法生成了简洁的结构化故事线集合，最大化地覆盖了显著的信息。**

还有一些其他的方法[5,6]考虑了路径级别的连贯性，从某种角度来说，它们汇总了链中所有文件之间的相似度值。然而，它们并未考虑文章的顺序或最弱环节的强度；尽管转变不佳，它们仍可能给链分配高连贯性。保证链之间的强转变有助于获取和理解知识。

在抽取文献摘要和可视化文献方面，存在多种工具；参见Borner[4]的纲要。与我们的文档集合不同，这些系统中有[12,16]很多使用了单一的概念作为分析单元。对于非专家来说，这种粒度太细，无法使用。粒度与我们相似的其他工具往往关注可视化引用或共引。[8,13]而且，文章之间的边基于局部计算，也没有连贯的研究线的概念。

最后，人们之前曾使用过与地铁图类似的视觉隐喻来展示抽象信息；例如，Nesbitt用地图展示了那些贯穿他博士论文的，相互关联的理念[21]。然而，这些地图均由人工构建，而我们的地图则是自动生成。

## 结论

我们概述了我们目前对抽取信息和构建汇总地铁图所使用的方法的研究。给定某个查询后，我们的算法生成了简洁的结构化故事线集合，最大化地覆盖了显著的信息。最重要的是，地铁图明确地展示了线条之间的关系。我们应用了地铁图帮助人们理解新闻故事、研究领域、司法案例和文学作品。我们在几个领域中利用真实的数据集进行了前景广阔的探索性用户研究。我们的结果说明，地铁图能够帮助用户更有效地获取知识。

我们的研究也存在若干局限，将来处理这些局限会是很有趣的事情；例如，我们的连贯性概念假定了词语的重复，所以我们的系统无

法处理非常短的文章（比如Twitter的博文）。不仅如此，我们的特性缺乏深度，使得连贯性度量偏向来源相同的文章链。我们还计划让我们的系统对噪声更鲁棒；虽然我们的方法对从数据集中移除一些文章的行为并不敏感，但是近乎重复的文章会影响覆盖面的权重，让算法偏向于覆盖这些文章。对查询日期范围中的微小变化更敏感这一情况或许可以通过自动找出时间线的最佳分段来处理。

我们还计划试验形式更丰富的输入、输出和交互机制，并整合较高层级的语义特性。当前，我们把很多工作投入到了设计目标函数上；将来，我们希望直接通过用户反馈学习或修正目标函数。另一个有趣的方向是观点机制，让用户通过他人的眼睛观察一个主题（比如民主党人询问共和党人的观点）。

这一研究方向可能会引出一些工具，帮助人们在信息爆炸的情况下浏览和理解理念、趋势、联系以及故事线。

## 鸣谢

我们在此感谢Rok Sosic、Andrej Krevl、Dima Brezhnev、Caroline Suen、Jeff Jacobs、Heidi Wang、Thomas von der Ohe、Tom Camenzind、Rohan Puttagunta和Raiyan Khan。本研究的部分资助来源于NSF IIS-1016909、CNS-1010921、IIS-1149837、DARPA SMISC、DARPA GRAPHS、ARL AHPCRC、Okawa Foundation、Docomo、Boeing、Volkswagen、Intel、布朗媒体创新研究所（the Brown Institute for Media Innovation）和阿尔弗雷德·斯隆奖（the Alfred P. Sloan Fellowship）。

参考资料

1. Ahmed, A., Ho, Q., Eisenstein, J., Xing, E., Smola, A.J., and Teo, C.H. Unified analysis of streaming news.In *Proceedings of the 20th International Conference on the World Wide Web* (Hyderabad, India, Mar. 28–Apr. 1).ACM Press, New York, 2011.
2. Allan, J., Gupta, R., and Khandelwal, V. Temporal summaries of new topics.In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, LA, Sept. 9–13).ACM Press, New York, 2001.
3. Bloom, B.S., Engelhart, M.D., Furst, E.J., and Hill, W.H., Eds. *Taxonomy of Educational Objectives, Handbook 1:Cognitive Domain*.Longman, White Plains, NY, 1956.
4. Borner.K. *Atlas of Science:Visualizing What We Know*. MIT Press, Cambridge, MA, 2010.
5. Boyack, K.W. and Klavans, R. Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology 65*, 4 (Apr. 2014), 670–685.
6. Braam, R.R., Moed, H.F., and Van Raan, A.F. Mapping of science by combined co-citation and word analysis, I:Structural aspects.*Journal of the Association for Information Science and Technology 42*, 4 (May 1991), 233–251.
7. Broder, A. A taxonomy of Web search.*ACM SIGIR Forum 36*, 2 (Sept. 2002), 3–10.
8. Chen, C. Citespace II:Detecting and visualizing emerging trends and transient patterns in scientific literature.*Journal of the Association for Information Science and Technology 57*, 3 (Feb. 2006), 359–377.
9. Druck, G., Mann, G., and McCallum, A. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, July 20–24).ACM Press, New York, 2008.
10. Faloutsos, C., McCurley, K.S., and Tomkins, A. Fast discovery of connection subgraphs.In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Seattle, WA, Aug. 22–25).ACM Press, New York, 2004.
11. Farrand, P., Hussain, F., and Hennessy, E. The efficacy of the 'mind map' study technique.*Medical Education 36*, 5 (May 2002), 426–431.
12 Fox, E.A., Neves, F.D., Yu, X., Shen, R., Kim, S., and Fan, W. Exploring the computing literature with visualization and stepping stones and pathways.*Commun.ACM 49*, 4 (Apr. 2006), 52–58.
13. Garfield, E. and Pudovkin, A.I. The histcite system for mapping and bibliometric analysis of the output of searches using the ISI Web of Knowledge. In *Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology* (Providence, RI, Nov. 12–17).Association for Information Science and Technology, Silver Spring, MD, 2004.
14. Gillenwater, J., Kulesza, A., and Taskar, B. Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island, Korea, July 12–14).Association for Computational Linguistics, Stroudsburg, PA, 2012, 710–720.
15. Hall, R.H. and O'Donnell, A. Cognitive and affective outcomes of learning from knowledge maps. *Contemporary Educational Psychology 21*, 1 (Jan. 1996), 94–101.
16. Hossain, M.S., Gresock, J., Edmonds, Y., Helm, R., Potts, M., and Ramakrishnan, N. Connecting the dots between PubMed abstracts.*PloS One 7*, 1 (Jan. 2012), e29509.
17. Jo, Y., Hopcroft, J.E., and Lagoze, C. The web of topics: discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th International Conference on the World Wide Web* (Hyderabad, India, Mar. 28–Apr. 1).ACM Press, New York, 2011, 257–266.
18. Kleinberg, J. Bursty and hierarchical structure in streams.*Data Mining and Knowledge Discovery 7*, 4 (Oct. 2003), 373–397.
19. Nallapati, R., Feng, A., Peng, F., and Allan, J. Event threading within news topics.In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management* (Washington, D.C., Nov. 8–13).ACM Press, New York, 2004, 446–453.
20. Nenkova, A. and McKeown, K. A survey of text summarization techniques.In *Mining Text Data*, C.C. Aggarwal and C. Zhai, Eds. Springer, 2012, 43–76.
21. Nesbitt, K. Getting to more abstract places using the metro map metaphor.In *Proceedings of the Eighth International Conference on Information Visualisation* (London, U.K., July 14–16).IEEE, 2004, 488–493.
22. Radev, D., Otterbacher, J., Winkel, A., and Blair-Goldensohn, S. Newsinessence:Summarizing online news topics.*Commun.ACM 48*, 10 (Oct. 2005), 95–98.
23. Rewey, K.L., Dansereau, D.F., and Peel, J.L. Knowledge maps and information processing strategies. *Contemporary Educational Psychology 16*, 3 (July 1991), 203–214.
24. Shahaf, D. and Guestrin, C. Connecting the dots between news articles.In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, D.C., July 25–28).ACM Press, New York, 2010.
25. Shahaf, D., Guestrin, C., and Horvitz, E. Metro maps of science.In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Beijing, China, Aug. 12–16).ACM Press, New York, 2012.
26. Shahaf, D., Guestrin, C., and Horvitz, E. Trains of thought:Generating information maps.In *Proceedings of the 21st International Conference on the World Wide Web* (Lyon, France, Apr. 16–20).ACM Press, New York, 2012, 899–908.
27. Shahaf, D., Yang, J., Suen, C., Jacobs, J., Wang, H., and Leskovec, J. Information cartography:Creating zoomable, large-scale maps of information.In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, IL, Aug. 11–14).ACM Press, New York, 2013, 1097–1105.
28. Swan, R. and Jensen, D. TimeMines:Constructing timelines with statistical models of word usage.In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Boston, MA, Aug. 20–23).ACM Press, New York, 2000.
29. Yan, R., Wan, X., Otterbacher, J., Kong, L., Li, X., and Zhang, Y. Evolutionary timeline summarization:A balanced optimization framework via iterative substitution.In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China, July 24–28). ACM Press, New York, 2011, 745–754.
30. Yang, Y., Ault, T., Pierce, T., and Lattimer, C. Improving text categorization methods for event tracking.In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece, July 24–28). ACM Press, New York, 2000, 65–72.
31. Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., and Liu, X. Learning approaches for detecting and tracking news events.*IEEE Intelligent Systems 14*, 4 (July/Aug. 1999), 32–43.

Dafna Shahaf（dshahaf@cs.stanford.edu）是加利福利亚州斯坦福大学计算机科学系博士后研究员。

Carlos Guestrin（guestrin@cs.washington.edu）是华盛顿州西雅图市华盛顿大学计算机科学和工程系副教授。

Eric Horvitz（horvitz@microsoft.com）是华盛顿州雷德蒙德微软研究院杰出科学家和主任。

Jure Leskovec（jure@cs.stanford.edu）是加利福利亚州斯坦福大学计算机科学系助理教授。

译文责任编辑：崔斌



观看作者在独家《通讯Communications》（视频中讨论他们的研究：http://cacm.acm.org/videos/information-cartography