DOI:10.1145/2629572

即便你无法确定你要找的信息，你仍可使用这个地图查询界面来搜索世界。

HANAN SAMET，JAGAN SANKARANARAYANAN，MICHAEL D. LIEBERMAN，MARCO D. ADELFIO，BRENDAN C. FRUIN，JACK M. LOTKOWSKI, DANIELE PANOZZO，JON SPERLING， BENJAMIN E. TEITLER

利用空间同义词在地图上阅读新闻

你旅行吗？你想知道在此次旅行目的地及其附近正发生的事件吗？你想了解你离开的地方及其附近最近的新闻吗，特别是你曾居住或工作过的地方？如果你对上述任一问题给出了肯定的回答，那么我们的报亭（NewsStand）【指新闻的空间-文本聚合及其显示（Spatio-Textual Aggregation of News and Display）】应用和相关系统恰好符合你的需要。

报亭[46]是支持人们利用地图查询界面来检索信息的通用框架的一个示例应用。如上所述，在之前的30多年里，我们一直在马里兰大学开发我们称为“空间浏览器”的系统，而上述应用则是其中的一个变体（见Samet等人的两篇论文[39，41]）。地图查询界面的优点是，由于结合了查看时缩放尺度可变的能力，地图可为搜索过程提供内在的粒度，加快近似搜索。这种能力使之与广为流行的基于关键词的常规搜索方法区分开来。常规方法提供的近似搜索功能有限，其主要通过匹配关键词的子集实现。然而，用户通常无法确定应该使用哪个关键词，因此他们希望搜索时能考虑同义词。在名为“对空间数据的空间查询（spatial queries to spatial data）”的空间参照数据的查询方面，地图查询界面取得了一定的进展。我们考虑了指向位置的行为（例如通过定点设备进行妥善定位或采用适当的手势），并依据缩放比例解释该定位规格的精度。这等价于允许使用空间同义词。

支持使用空间同义词相当重要，因为它支持用户在不确定查询目标时或不确定查询应该返回的结果时搜索数据。例如，假设用户查询是“在曼哈顿举办的摇滚音乐会（rock concert in Manhattan）”。如果找不到在曼哈顿举办的此类事件，那么在哈莱姆、布鲁克林区或纽约市举办的摇滚音乐会都算是相当匹配的答案，因为它们对应于曼哈顿的空间同义词：哈莱姆在曼哈顿区之

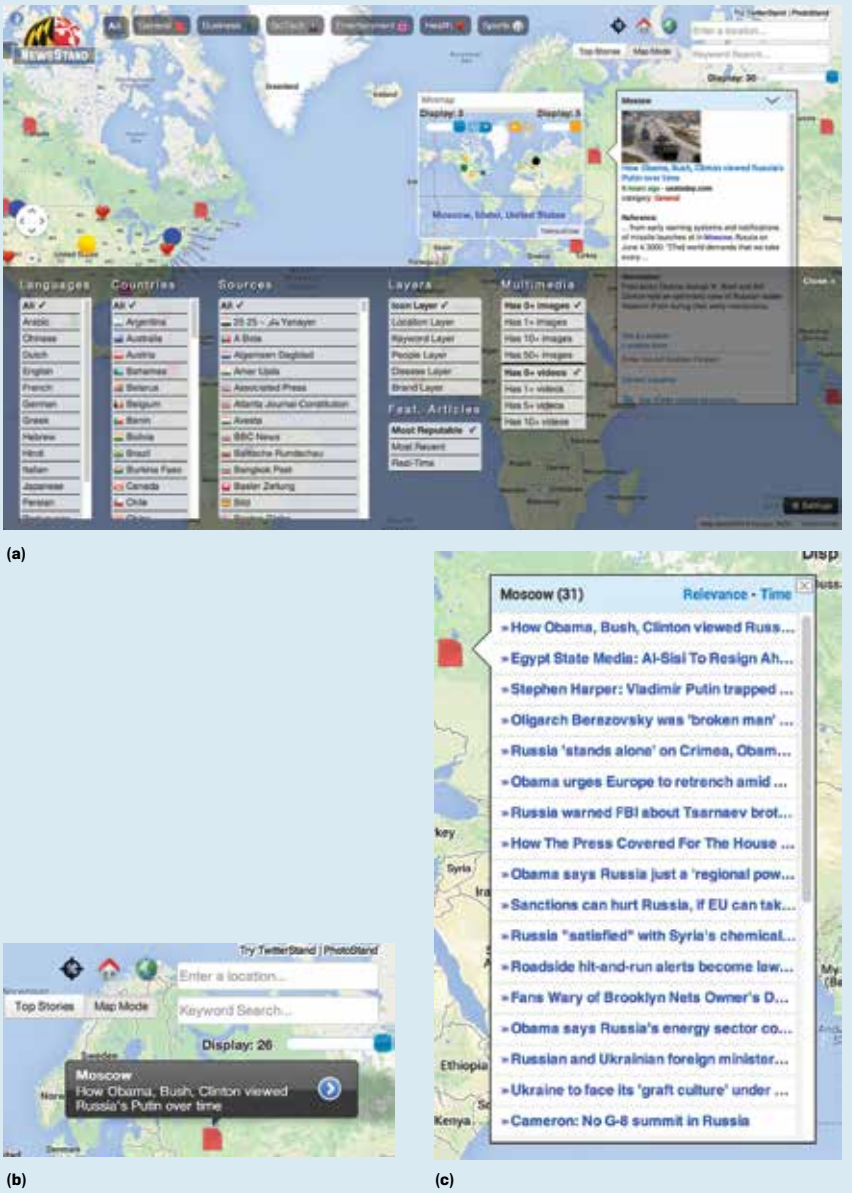
» 重要见解

- **报亭的地图查询界面监视着10,000多个发布时间间隔在数分钟之内的RSS新闻源的输出，并把文章与文中提到的地点关联。**
- **在结合查看和解释时缩放尺度可变的能力后，地图可为搜索过程提供内在的粒度，加快近似搜索，并允许使用空间同义词。**
- **对于移动设备的用户来说，地点的文本规范虽较几何规范更好，但却必须解决潜在的模糊性。**

插图由COHERENT IMAGES提供


BREAKING NEWS
Green Knoll Golf and Tennis
Best Buy
Commons Way
Somerset Corporate
Bond St
Birdsall Ln
Birdsall Ln
Country Club Rd
Golf Links Dr
Alletra Ave
206
202
22
28
287
Easton Turnpike

图 1.报亭的地图模式：（a）“2014年3月26日地点X处正在发生什么？”；（b）奥巴马/普京关系主题的代表性标题；以及（c）与莫斯科相关联的主题的代表性标题。



内；布鲁克林区则靠近该区且与该区的等级相同（两者均为纽约市下面的区）；而纽约市则包含了该区。常规的搜索引擎通过动态纳入从搜索 - 点击日志中收集的信息处理空间查询；据此，如果有足够多的用户在搜索曼哈顿时最终点击了与哈莱姆或纽约有关的网页，那么随着时间的推移，搜索引擎会推断出文档的空间范围最接近纽约，或与纽约相关。最近，搜索引擎（如谷歌的知识图和微软的 Satori）正在使用大型的知识库来理解关键词搜索查询的空间焦点，以及在一定程度上理解文档的空间焦点。尽管搜索引擎在理解文档的地点方面已经取得了上述进展，但是搜索引擎的主要工作原理仍然基于流行度。从这个角度来说，网页排名（PageRank）算法和点击日志确保向用户提供的网页会通过考虑了某种频率因素的测量值进行排序。具体而言，经典的网页排名算法使用了静态数据，但点击日志对应了动态数据。基于频率的方法确保其向某用户提供的结果与其他用户的结果相同。这种性质可以被刻画为“搜索的民主化”，就是说所有用户得到了相同的待遇。用更直白的方式来说，产生的效果是对用户不加区分，就是说他们得到相同的坏（或好）答案。换言之，使用网页排名算法和点击日志来对结果进行排序的效果（高效地选择向用户呈现的结果）是，如果没有人在之前曾搜索过某些数据（或空间意义上的近邻）或链接到该数据上，那么该数据将永远无法找到，因此也永远不会呈现给用户。在某些情况下，这种方法是可行的。然而，就同义词而言，这种方法对搜索结果的质量施加了相当强的负面影响，因为这意味着，对于不使用相同词语但内容等同的页面，倘若没有人链接到该页面，或点击空间邻区，那么搜索引擎将永远无法找到相似性。同样地，网页排名算法也永远不能找到类似的页面；在构建网页索引时它会在网络上抓取信息，但它却找不到有用的点击日志。

我们在马里兰大学中搭建的报亭和相关系统处理了空间查询中的同义词问题。注意，所有的空间查询均可归为两类：

*基于地点的查询*给出地点 X，传统上使用经纬度坐标值作为参数，并返回与 X 相关联的一组特征集合；以及

*基于特征的查询*利用特征 Y 作为参数，返回与 Y 相关联的一组地点集合。

这些查询还可用两个函数来刻画，其中一个函数为另一个函数的反函数。基于特征的查询也被称为“空间数据挖掘”。[3] 虽然特征通常为空间参照数据（例如作物类型、土壤类型、地带和速度限制）的特性（或称为属性），但是它们及其底层的空间参照数据域的解释可能更宽。报亭把它们转换成由新闻文章的集合组成的非结构化数据域，

其中新闻的地点通过文字标明；而各种特征则为主题。转换这些概念后，基于地点的查询会返回提及特定地点或区域 X 的所有主题和文章，而基于特征的查询则返回与主题 T 有关的文章或只在文章 Y 中提及的所有位置和区域。注意，报亭不需要用户提前指定 T。如果未指定，则主题会按重要程度排序，其中的重要程度可通过多种标准定义，包括但不限于包含的文章数量。下文举出了两个典型的查询示例：位置 X 处正在发生什么？主题 T 或文章 Y 的发生地点在哪儿？

它们的执行通过构建空间数据的索引来加速，最好在批量加载的过程中立即构建所有的索引[36]，如 Hjaltason 和 Samet 的工作[12]所述。在当空间数据拥有确切的地理信息和数值信息时，构建索引相对容易。然而，由于所有数据都是非结构化的，所以报亭中数据的描述方式并非如此。具体来说，位置和特征数据都只是词语的集合。从空间数据的角度来看，其中的某些数据可以解释（但并不要求这么做）为地点的名称。换言之，空间数据通过文本（称为“地名”）而不是几何数据描述，也就是说会有一些模糊性。这种模糊性既有好处，也有坏处。好处是，从几何的角度来看，文本规格从点和空间范围两方面解释了数据，这点与参数传递中的多态类型相似（它构成了面向对象编程语言中继承的基石）。例如，在地理上，某个城市可以用某个点（比如它的形心）或与其边界对应的区域来确定。选择何种方式取决于所激活的查询界面的缩放尺度。坏处是，我们无法确定搜索项是否一直是地理位置。例如，在“Michael Jordan”中，“Jordan”是指国家，河流，还是姓？上述解答过程称为“地名识别”。[18]不仅如此，如果它是地理位置，那么如果有很多名称相同的地理位置实例，它指哪个位置。例如，“伦敦”指英国伦敦市、加拿大安大略省伦敦市，还是其他的地方？上述解答过程称为“地名分辨”。[19]在部署报亭和相关系统时，如何正确无误（或基本无误）地弄清这些模糊点是我们面临的主要技术挑战。

报亭的用户界面

报亭的目标不仅是提供一种不同的新闻阅读过程，更重要的是另一种体验。在报亭中进行查询时，用户需要选择感兴趣的区域，然后找到相关的关联主题和文章（请访问 http://newsstand.umiacs.umd.edu 体验报亭的界面）。主题和文章的展示由地点和缩放尺度确定，它们共同决定了查询的空间范围，或者说感兴趣的区域。“感兴趣的区域”这一概念有两种解释方式，一种从内容角度，一种从新闻源角度。用最简单的方式来说，对于感兴趣的区域，系统展现了相关的文章，但并没有预先限制发表这些文章的新闻源的地点范围。其次，通过明确指定新闻源（比如《*纽约时报*》和《*华盛顿邮报*》）、语言、用文本方式指定空间区域（比如限定新闻源的范围为爱尔兰）、或在报亭的地图上圈出感兴趣的区域（比如覆盖爱尔兰和英国的方框），可以把新闻源限定在可用新闻源的一个子集内。用户也可以约束空间区域和新闻源；它们不一定要相同。这个功能相当有用，因为它允许用户了解世界的某个区域如何看待另一个区域发生的事情。例如，用户可能希望了解英文媒体如何看待和解读中东的动态。这种结果类似于情感分

图 2.报亭的头条模式：（a）查询“2014年3月26日主题 T 或文章 Y 正在何处发生？”时的截屏示例；以及（b）与奥巴马/普京关系主题相关联的图片的子集，其中重复和近似重复的图片已经置灰。

(a) (b)

析。其他的应用包括为投资者监控热点地区、国家安全以及得到疾病扩散的最新消息，参见 Lieberman 等人的论文。[24]

图 1a 为查询“2014 年 3 月 26 日地点 *X* 处正在发生的事件（What is happening at location X on March 26, 2014?）”时报亭输出的截屏。该图中使用了报亭的“地图模式”。*X* 是非洲、欧洲和美洲部分地区。图中包括了一则有关奥巴马 / 普京的文章片段，其中论及了莫斯科。我们把地图上的每个图标或标识称为“标记”，其代表了一组主题相同和 / 或不同的文章集合，其中所有文章共有的主要性质为它们均提及了对应的地图位置。标识的类型说明了新闻类别（比如普通新闻、商业、科技、娱乐、健康和体育），范围涵盖了与该位置关联的大多数文章主题。用户可以切换屏幕顶部的相应按钮状态选择其中的一个或多个类别。

图 1b 为一个信息提示框，包含了与莫斯科或奥巴马 / 普京关系相关的主要主题中的代表性文章。报亭通过对所有文章使用聚类过程获取这些主题。当用户把鼠标停在莫斯科上时，系统生成了该信息提示框。这种停留的行为还会让地图上与该代表性的文章相关联的所有其他地点的标记变成橘球。在本例中，这些地点部分与奥巴马 / 普京关系涉及的或受其影响的国家对应。某些地点没有出现在截图中，可能是因为它们处于当前可见的地图的地理范围之外（比如北美和远东）。

当用户把鼠标放在标记上时，系统会生成迷你地图和标题（未在此处展示），其中会包含大地图之外的感兴趣区域。用户的这种行为会让橘球会出现在适宜的位置上，展现了代表性文章的地理范围。迷你地图这一工具允许用户查看所选文章的地理焦点，同时不必离开主地图上的感兴趣区域。而且，它与当前的缩放尺度无关。

报亭通过抓取网页收集数据。它的主要数据源为世界上以真正简易聚合（RSS）订阅源的形式存在的成千上万个单独的新闻源。

主地图和迷你地图上的蓝球说明了与用户鼠标当前停留的位置（此处为莫斯科）名称相同的其他地点。在允许迷你地图中包含名称相同的其他地点后，迷你地图的地理范围可能会超出橘球的范围。蓝球支持检测地名分辨错误。

迷你地图上的黑球标出了用户鼠标当前停留的地点，即莫斯科。迷你地图上的上下箭头允许用户滚动查看橘球和蓝球，并输出对应的地点名称。滚动查看蓝球时，系统支持对地点名称的解释进行排序。迷你地图上的绿球和红球对应滚动过程中的当前蓝球和橘球。把鼠标放在迷你地图上的橘球上时，会出现地点的名称；而放在蓝球上时，因为所有的蓝球名称相同，系统会在迷你地图上同时显示该地点及其上级地点的名称（如“Moscow（莫斯科），ID，United States（美国）”）。

图 1c 中的信息提示框展现了代表性文章的标题，这些代表性文章属于与莫斯科关联的各主题，而莫斯科则是用户鼠标最近停留的位置。单击与这一地点关联的标题信息提示框中的 > 标记后，便可得到该图。单击标题列表中的某一标题后，系统会展示汇总信息提示框（见图 1a），并在旁边展示对应的迷你地图。它也是在用户把鼠标放在标记上后生成的。注意，当用户滚动查看主题中的标题时，迷你地图上的橘球（不是蓝球）会发生变化。汇总信息提示框还包含了相关图片、视频和其他文章的链接。单击汇总信息提示框上的标题后，系统会展现文章的全文。而且，如果语言不是英语，还会有一个通过翻译包（如谷歌翻译或微软翻译）把全文和 / 或标题翻译成英文的选项。

新闻源的域可通过语言、地理区域或国家以及特定的报纸加以限制，代表性的文章从这些源的文章

中抽取。该功能通过使用“settings（设置）”按钮（位于图 1a 中屏幕右下的角落）配置和选择合适的过滤器实现，如图 1a 中靠下的灰色部分所示。注意，用户还能通过地点或（多个）关键词进行搜索，并通过控制显示的滑动条改变展现的标记数量。

图 2a 为查询 2014 年 3 月 26 日“主题 *T* 或文章 *Y* 正在何处发生？（Where is topic T or article Y happening on March 26, 2014）”时，报亭输出的截屏。这是报亭的“头条模式”。*T* 为主题之一，其中的代表性标题展示在左侧底部的面板内，按重要性度量排序。重要性按照显著性、距现在的时间和频率定义。虽然本应考虑主题到达的速度 / 加速度，这是个更好的度量标准，因为主题最终会失去其时效性。显示的标题为曾经点击过的标题。当用户鼠标放在上面时，它会被突出显示（通过变灰），此处对应的是奥巴马 / 普京关系的主题。单击标题后，会出现与标题相关的详情（比如更为详细的描述，相关文档、图片和视频的数量），如图 2a 左侧顶部的面板所示。同时，还可通过继续点击鼠标访问这些内容。

把鼠标放在图 2a 左侧底部面板上并点击后，系统还会在地图（右侧面板）上与主题关联的主要地理位置处展示合适的标识（类别标记）。在本例中，这些地点部分对应与涉及奥巴马 / 普京关系，或受其影响的国家，包括美国和俄罗斯。把鼠标放在右侧面板的地图上后，会出现很多信息提示框和关联的迷你地图，其中迷你地图与提示框的语义相同，如图 1 中“*X* 处正在发生什么？（What is happening at X）”的查询所示。具体来说，橘球支持用户区分临近的多个地点（比如网球簇中的英国伦敦和温布尔顿），而蓝球则标出了名称相同的其他地理位置实例【如“美国宾夕法尼亚州莫斯科镇（Moscow, PA, United States）”】。

使用地图和头条模式后，用户可以获得与各簇关联的图片和视频集合。而且，报亭会检测出重复或近似重复的图片，在视图中予以隐藏。这是一种功能强大的特性；首先，系统使用了与文章相关联的词或其语义找出相似的图片，尔后通过经典的图像相似性方法（包括分层颜色直方图[5]和尺度不变特征转换算法，即 SIFT）检测出相似图像中的重复。[25] 图 2b 说明了锚定在莫斯科与奥巴马 / 普京关系主题关联的此类图像的子集。

正如之前论述的那样，报亭的最终目的是把地图作为展现与空间相关的信息的媒介，因此它不限于新闻文章；也就是说，它还可用于搜索结果、图片、视频和推文。它还支持新闻汇总，深层探索，甚至通过发现新闻中的模式进行知识获取。把主题和类别与组成的文章中所提及的地点相关联后，便可直接得到这一结果。例如，查询可以被链接在一起。从这个意义上讲，有趣的主题可能与法国巴黎相关联，但在浏览橘球时，相同的主题或许又会与英国伦敦相关联。此时，用户可把定点设备移到伦敦上面，单击后，可以找到提及伦敦的其他相关主题和其他地点，然后用户又可以通过在地图查询界面上移动而转换到那些地点。这种无限的链接只在地图模式下提供，因为此时查询是基于地点的；在头条模式下，因为查询是基于主题的，所以除非用户使用了关键词搜索，否则地图上的标记限于排名最高的主题所对应的地点。

报亭还支持计算发病中心的簇，即簇中与疾病名称对应的最常用搜索项【比如图 3 中“2014 年 3 月 26 日的欧洲（Europe on March 26, 2014）】。另外，用户也可使用相同的理念，在簇中找出与人名或品牌名对应的最常用搜索项。在分别把“层”的参数设置为“疾病”，“人”或“品牌”后，便可找到此类搜索项。

相关研究

很难把报亭与现有的新闻阅读器进行比较，因为所有流行的新闻阅读

图 3. 报亭截屏展示了2014年3月26日欧洲各国中提到某疾病名称的簇；用户把鼠标放在西班牙巴伦西亚上，疾病名称为乳腺癌。迷你地图上的橘球展现了世界上的其他地点，与那些地点关联的簇提到了乳腺癌。

器（比如 Pulse）尚未具有使用地图阅读新闻的功能。新闻阅读系统（比如微软必应新闻、谷歌新闻和雅虎新闻）用经典的线性方式展现新闻，把来自不同源的新闻按主题归类。从根据用户所在位置聚合相关文章和主题的角度来看，这些提供者均含有某些位置特征。聚合通常依照邮编或市 - 州规定完成。例如，对于邮编 20742，主题可能提到了“马里兰州帕克分校”。谷歌新闻好像实现了这种功能。至少据我们所知，在谷歌搜索中使用地名作为搜索词可以得到类似结果。例如，确定用户的邮编为 20742 后（比如，在缺少获得用户所在当地区域的其他规定时，使用用户的 IP 地址），谷歌新闻会返回提及“马里兰州帕克分校”或“马里兰大学”的主题，因为已知它们与该邮编相关联。另外，主题的结果列表主要基于新闻源（通常为报纸）的地点，而不是报道的内容。新闻源包含了组成主题的多篇文章。在上述示例中，展现的主题数量是有限的。除了排除与用户位置无关的主题外，使用这种限制并无其他特别的原因。还请注意，在这些示例中，在决定向用户展现的内容时，没有使用文章重要性的概念。

有趣的是，流行的新闻阅读器均没有使用地图展现文章，虽然它们只要在地图平台上采用糅合（mashup）便能实现这一功能。HealthMap[10] 确实使用地图来展现疾病的暴发，其中的地点从疾病报告的日期栏或 ProMed 报告的元数据处获取。使用地图展现疾病报告与报亭的“疾病层”功能相似（见图 3），不过报亭中的地点来源于文章的实际文本。它还与我们在抽取网络上的空间 - 文本助力文档检索（STEWARD）系统中的实现方式相似 [23]。该系统利用了 ProMed 报告，也可展示疾病随时间的传播情况。[16] 注意，虽然支持糅合（mashup）的地图平台能够放大，但除了报亭之外，尚没有地图平台把缩放与获取更多文章的能力相结合。

过去，有些系统试图理解和展现新闻文章中的地理位置，但其中的大多数已经无法找到，或无法访问。例如，路透社的 NewsMap、*华盛顿邮报*的 TimeSpace、BBC 的 LiveStats 以及 AP 的 Mobile News Network（移动新闻网络）均试图根据提交文章的新闻通讯社的地点把新闻文章与大致的地理位置关联起来。因此，向迈阿密新闻通讯社提交的文章将会与迈阿密所有的邮编关联。与报亭不同，AP Mobile News Network 似乎不打算分析单独的文章，进而确定主要的关联地点，或地理焦点，或文章中提到的其他重要地点。

把报亭和网络搜索及推荐系统的商业服务（比如评论站点 Yelp 和 TripAdvisor）进行比较后，我们也得出了一些有用的信息。区别在于，在那些系统中，在感知空间实体前，需要在系统的数据库中明确输入以地址或 GPS，或经纬度值记录的空间信息；因此，它们能支持对空间信息的探索。报亭则起着两种作用：发现输入数据中的空间信息，这些信息通过文本描述，通常具有模糊性（需要纳入其他信息，有些未在输入数据内）；以及探索性的作用，其中的功能与推荐系统的功能相似，虽然推荐系统对地图查询界面的重视程度较弱。

报亭的架构

在 Rudyard Kipling 于 1902 年出版的《*原来如此的故事（Just So Stories）*》中，他对理解新闻所需的关键因素做出了或许是最好的阐释：“我雇佣了六个诚实的仆人（他们教会了我所有我知道的事情）；他们的名字是什么、哪儿、什么时候、怎么样、为什么和谁（What，Where，When，How，Why，Who）。”报亭关注于“什么”和“哪儿”，但较少关注“什么时候”，“什么时候”通常指的最近的时间。此处，我们先关注“什么”，随后再关注“哪儿”。

报亭通过抓取网页收集数据。它的主要数据源为世界上以真正简易聚合（RSS）订阅源的形式存在的成千上万个单独的新闻源；RSS 是在线出版中广泛使用的 XML 协议，非常适于报亭，因为它只需要一个标题、简短描述以及每则已刊发新闻的网络链接。RSS 2.0 还允许加入可选的发布日期，这可帮助确定文章距现在的时间，或者说“新近程度”。报亭现在为 10,000 个新闻源构建了索引，每天处理约 50,000 篇新闻文章。它使用名为地理标记的过程确定文章中提及的地理位置，而且设法确定文章的地理焦点或中心点，即文中提到的关键地点。

报亭还根据内容相似性把新闻文章按主题归类（称为“聚类”），所以与相同事件有关的文章会被归为相同的簇中。聚类的主要目的是自动对新闻文章进行分组，对描述相同新闻事件的文章进行归类，形成名为“文章簇”的新闻文章集合（之前文中也将其称为“主题”或“簇”）。之后，每个簇便只包含当前得到的输入中与特定主题相关的文章。新闻文章进入此阶段后，报亭会把它们归入新闻簇中。这本质上是一次性的过程，也就是说，一旦文章加入簇后，它会一直在簇中存在。报亭永远不会重新处理文章或重新对文章进行聚类。因为进入报亭的文章吞吐率很高，而且报亭的文档聚类系统需要能够快速处理文章且保持优质聚类输出，所以这点符合我们的期望。这种版本的聚类算法具有“在线”的特性。

给定上述需求后，报亭使用了领导者—追随者聚类 [7] 算法。该算法允许从使用词频——逆文档频率

（TF-IDF）的搜索项 - 向量空间度量和[35]时间维度两个方面进行在线聚类。对于每个簇，报亭维护了一个搜索项形心和时间形心，分别对应于簇中所有搜索项 - 特征向量的均值和文章的发布时间的均值。对新闻文章 a 进行聚类时，报亭检查是否存在搜索项与时间形心至 a 的距离小于固定截断距离 ϵ 的簇。如果存在一个或多个备选聚类，则把 a 加入最近的备选簇，并更新簇的形心；否则，报亭创造一个新的簇，其中只包含 a。

报亭的在线聚类算法依照其“重要性”的概念对簇进行排序。“重要性”由以下几个因素决定：

*文章的数量。*簇中文章的数量；

*簇中唯一的新闻源的数量。*例如，如果多家新闻源报道了在加利福尼亚州欧文市发生的事件，特别是如果有一些位于相隔较远的洛杉矶（距欧文市约 50 英里），则该事件属于重要事件。

*簇的传播速度。*有关重要事件的新闻会在较短的事件内被多家新闻源报道；以及

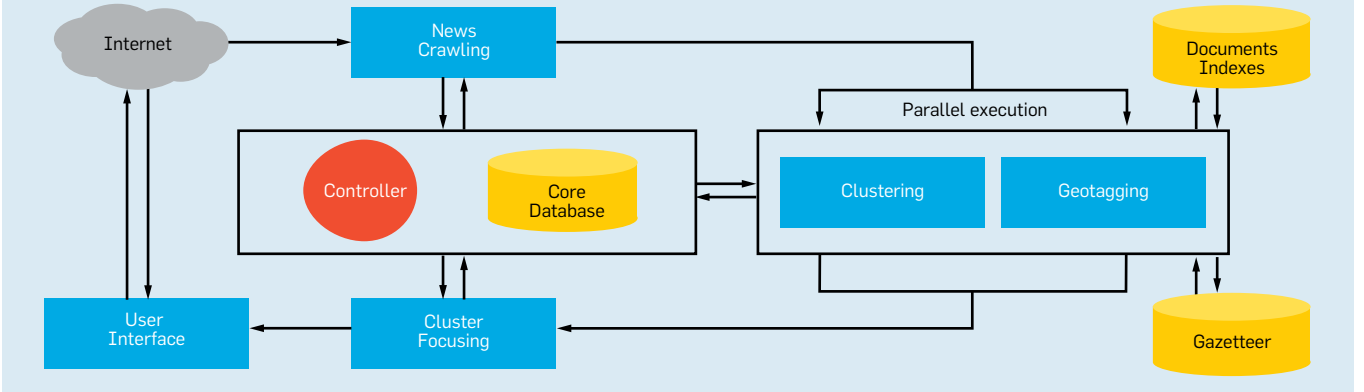
*添加的时间。*最近加入簇的时间。这是报亭用户可以设置的选项，可让系统忽略前三个因素。

当使用前三个因素对簇进行排名时，报亭必须选择簇的代表性文章，这是一种二级排序。该文章的性质可能因报亭用户的设置不同而不同，或是最近的文章，忽略对应簇的重要性（第四个因素），或是根据簇的重要性进行筛选；其中的选择范围或是声誉最好的新闻源的文章，或是最新的文章。在地图模式下，报亭在当前的观察窗口中展现的簇须含有最显著的主题，这点相当重要；仅仅在地图上展现排名最高的主题可能无法给广大的受众提供有用的结果，因为这些主题倾向于聚集在特定的地理区域中。这种情况反映了大型报社的新闻覆盖不均的情况，因为它们倾向于关注自己的地理区域。在报亭中，主题选择是在显著性和范围之间进行权衡。为了达到平衡，报亭把展现窗口细分为规则网格，要求每个网格区域内包含的主题数不多于最大主题数。展现的主题按显著性和距现在的时间降序选择，该方法确保热门主题在整个地图范围内良好分布。

报亭也能测定与簇关联的地理焦点或中心点；通过与地点特征对应的聚类过程，系统加快了这一测定过程。只要某个簇是最重要的簇之一，或其地理焦点与最重要的簇之一的焦点相同，则报亭会在该簇的地理中心点的位置处展现该簇，其中地点的数量通过地图右上角的滑动条调节。因此，与最重要的簇相关联的位置也是地图中含有数据的位置。这种展现通常会利用与新闻类别相对应的标识，如图 1 所示。然而，除了展现与簇相关联的类别标识之外，通过要求用户设置合适的“层”参数（如图 1a 所示），报亭还能展现与簇中最流行的搜索项（我们称之为“关键词”）相对应的文本。另外，通过设置层参数为“地点”，用户还能查看作为地理焦点的地点的实际名称。

在设计报亭的架构时，最重要的准则是规模化和对单篇文章的快速处理[20]（见图 4）。其他的目标包括，尽快（网上发布后几分钟内）展现最新的新闻，以及健壮，不易崩溃。通过把采集和处理细分为多个模块，报亭的架构满足了这些准则。其中的每个模块均可在分布式计算机集群中的不同计算节点上单独运行。图中勾勒出了计算管道中一连串模块对文章的处理过程。因为每个模块可能在不同的节点上运行，给定的文章可能会在系统的多个不同计算节点上处理。设计模块时，我们还支持任何模块的多个实例在一个或多个节点上同时运行。因此，依照其收到的新闻数量，报亭能够启动足够多的模块实例来进行处理。每个模块接收输入，然

图 4.报亭架构的高级概述我们把系统设计成管道，其中各单独的处理模块独立运行。中心控制模块通过把工作分配到另一模块并跟踪管道中的文章实现文章处理工作的编排。



后把输入存入作为同步点的 PostgreSQL 数据库。用户在报亭界面上的动作（比如缩放、平移和选择）会自动转化成 SQL 查询语句，PostgreSQL 数据库会返回查询结果。

地理标记

报亭从新闻文章中抽取地理位置（称为“地理标记”），并与地理信息检索的成果息息相关。在该领域的现有成果中，有很多处理了如何找出网站和单个文件的地理范围这一问题。在新闻文章的场景中，报亭区分了三种类型的地理范围：[26]

提供者。出版者的地理位置；

内容。文章或主题内容的地理信息；以及

服务。基于读者的所在地点。

报亭依据文章的内容确定文章的地理范围，同时还设法使用已知的提供者的范围以及通过学习得出的服务范围。

报亭扩展了我们之前在 STEWARD 中取得的地理标记成果[23]，以支持对暗网中的文档进行空间 - 文本查询。虽然 STEWARD 技术可用于任意的文档集合，但是报亭包含了其他的模块和功能。为提高新闻文章的处理效率，我们专门设计了这些模块和功能。STEWARD 处理各文档时，通常独立处理单个文档而不考虑和所有其他的文档的相关性。而把通常源于多个新闻源的文档归类为主题簇后，报亭利用了与主题相关的多个文档版本和实例，这样便可改进地理标记，让用户更容易获取相关文章。

地理标记包含两个过程：地名识别和地名分辨。地名识别涉及地理 / 非地理的模糊性，其中一个给定的短语可能指一个地理位置，也可能指其他类型的实体（比如，提到“华盛顿”时，需要确定它是地点，还是其他实体，比如人名）。别名的使用是第二个问题，其中多个名称指向同一个地理位置【比如“洛杉矶（Los Angeles）”和“LA”】。地名分辨又称为“地理名称的模糊性”或多义现象，涉及地理 / 地理的模糊性，其中给定的名称可能指多个地理位置中的任何一个。例如，“斯普林菲尔德（Springfield）”是美国很多城市的名称，包括马萨诸塞州的斯普林菲尔德市和伊利诺伊州首府斯普林菲尔德市。

图 5.居住在俄亥俄州哥伦布市附近的读者所拥有的地方词库的示例；注意，很多地方的地名与其他地区中名气更大的地名相同。

地名识别。现已可采用很多不同的方法进行地名识别，不过所有的方法均有一些共同的特点。识别的理念是在给定周围场景的情况下抽取“有用的”短语，或者最有可能提及地理位置和其他实体的短语。这些短语统称为文章的“实体特征向量”（EFV）。确定 EFV 时，最容易的方法是查找文章中已经在地名词典或地名和地点数据库中存在的短语。很多研究人员把这种方法作为他们的主要研究手段。[2] 具体来说，把地理信息与网页相关联的 Web-a-Where[2] 系统使用了一个小型的、结构良好的地名词典，其中包含了约 40,000 个地点。这个词典是通过收集人口数大于 5,000 的乡村和城市的名称而创建的。词典的规模严重限制了 Web-a-Where 实用的地理标记能力，因为这使得它无法识别人口较少的（往往是地方上的）地点。而这些地点在来自地方新闻源的文章中很常见。不仅如此，规模不大的地名词典也意味着 Web-a-Where 更容易出现地名识别错误。与使用较大的地名词典相比，它错失了分清地理 / 非地理的模糊性的机会。

为了处理较大地名词典中内在的地理 / 非地理模糊性，包括 Martins 等人[27]、Rauch 等人[33] 和 Stokes 等人[45] 在内的研究人员已经提出了多种启发式方法用于过滤可能错误的地名。MetaCarta[33] 识别了空间线索词（比如“city of”），以及特定格式的邮寄地址和地理坐标的文本描述。然而，在进行新闻文章的地理标记时，这种策略会引起严重的问题，因为每篇文章中通常都会包含报社总部的地址。由于 MetaCarta 主要关注更大的显著地点，这些格式良好的地理字符串在其地理标记过程中的作用太大，导致了很多地理标记错误。

其他地名识别的方法则扎根于自然语言处理中相关问题的解决方

案。例如，命名实体识别（NER）[47] 关注名词和名词短语，旨在从文章中找出与各种实体类别（如 PERSON、ORGANIZATION、 和 LOCATION）相对应的名词短语。标记为 LOCATION 的短语是最有可能成为地点的短语，被保存为实体特征向量的地理特征，而 ORGANIZATION 和 PERSON 短语则被保存为非地理特征。NER 方法可大致归类为基于规则的方法 [18,31] 或基于统计的方法。[17]

基于规则的解决方案以规则目录为特征，其中列出了地名可能出现的多种场景。另一方面，基于统计的解决方案依赖于标注后的文档语料库，使用这些语料库来通过类似隐马尔科夫模型（HMM）[47] 和条件随机场（CRF）的构件训练语言模型。[15] 在可以获得标注后的语料库时，HMM 和 CRF 被广泛使用。报亭的地名识别过程使用了 LingPipe 工作包 [4] 中的 NER 标记器。该标记器根据消息理解会议（MUC-6）和知名的 Brown 语料库提供的新闻数据进行训练。[9]

注意，NER 标记并不排除使用地名词典。与此相反，这些标记方法可作为过滤器或剪枝策略，用以控制对地名词典的查询量。缺点是，如果实体未被确认为潜在的地点，则会漏过该地点。这种情况偶尔会发生。报亭使用了最初从 100 多个地名词典中整理出的开源地名词典 GeoNames（http://geonames.org/），其中包括 GEOnet Names Server（GEOnet 名称服务器）和地理名称信息系统（Geographic Names Information System）。它现由世界各地的志愿者维护，包含了约 850 万个不同的地理位置的名称，其中约 550 万个名称是唯一的，而其他的名称用于进行地名分辨或解决地理 / 地理模糊性。由于报亭需要跟踪多语言下每个地点的名称，报亭的地名词典中包含了约 1630 万个地名。

最近使用报亭处理八百万篇文章的过程中，我们只碰到了约 60,000 个不同的地点，但有 40,000 多个面临地理 / 地理模糊性的问题，这使得地名分辨变得至关重要。地名词典还包含了有人居住的位置或区域的人口数量以及层次信息，包括包含该地点的国家和行政区划信息。这些信息在识别范围相当小的，地方上的地名时有用。我们把地名词典查找应用于每个地理特征 $f \in EFV$ 和匹配地点，以生成集合 $L(f)$，其中集合的数量与特征或 $|EFV|$ 的数量相同。

地名分辨。识别地名后，报亭使用地名分辨程序解决地理 / 地理模糊性。地理 / 地理模糊性分辨存在的问题与另一个更普遍的问题有关，即如何关联规范实体与文档中提及的每个名词短语，其又被称为“命名实体消歧”（NED）。为了消除名词短语的歧义，NED 采用了利用知识库（比如维基百科、DBpedia 和 Yago）匹配名词短语的方法。进行高级处理时，文档中提及的名词短语首先与多个备选实体进行匹配，然后根据知识库中这些实体的关联度进行消歧。例如，Milne 和 Witten[29] 使用了具有相关性度量的有监督学习方法，其中两篇维基百科文章的关联度依据两者均包含的导入链接数确定。类似的，Hoffart 等人 [13] 使用各种备选实体之间的“连贯性”来区分所有的名词短语。最近的某些研究已经设法把 NER 和 NED 模块整合为一个命名实体识别和消歧（NERD）模块 [34]，该模块扫描文档，然后输出其中提到的实体。

最简单的地名分辨策略是使用某种显著性度量（如人口）为每个识别出的地名分配一个默认的意义。包括 Amitay 等人，[2] Martins，[27] Purves 等人，[31] Rauch 等人，[33] 和 Stokes 等人 [45] 在内的很多研究人员已经结合其他的方法实现了这种策略。例如，根据给定地名的每种解释在语料库中出现的频率，MetaCarta[33] 用概率的形式为分配了“默认意义”。这种语料库由预先采集的有地理标注信息的文档组成。然后，它会根据其他的启发式方法（比如线索词和邻近地名的出现次数）改变这些概率。互联网空间感知信息检索（SPIRIT）项目 [31] 使用了与 MetaCarta 相似的技术。它查找了句子线索，在没有更强证据时，会为给定的地理参照分配“默认的意义”。

注意，使用基于语料库的默认意义和概率后，系统几乎无法识别文章中相对来说没有名气的地点参照（比如，世界上 2,000 多个名气较小的“伦敦”实例中的任意一个）。这需要选择当地报纸的文章作为正确的解释，因为预先创建的新闻文章语料库中极少会出现这些名气较小的地点。相比之下，报亭使用了我们称之为“地方词库”[22,32] 的概念。该词库与新闻源进行了关联，包含了位于新闻源的地理范围之内的地点集合。例如，居住于“俄亥俄州哥伦布市”的读者的地方词库包含了“都柏林”、“阿姆斯特丹”、“伦敦”、“特拉华”、“非洲（Africa）市 "、“亚历山大”、“巴尔的摩”和“不来梅”（见图 5）。对于哥伦布市区域以外的读者，由于其地方词库中没有包含这些位置名称，所以在名称相同时，他们可能先考虑更有名的地点。

使用地方词库与之前描述的使用提供者 - 和服务 - 范围对地理范围进行解释的情况类似。具体来说，报亭通过构建每个新闻源的文章语料库和收集语料库中提及的地方地理位置信息来学习自己的服务范围。该方法基于以下观察结果，书写新闻文章时假定了读者的位置。例如，当伊利诺斯州（比如芝加哥）的报刊文章提到地点“伊利诺斯州斯普林菲尔德市”时，限定词“伊

利诺斯州”或“IL”很有可能不会出现，因为读者可以自动做出正确的解释。另一方面，在讨论“斯普林菲尔德市”时，《*纽约时报*》则需要保留“伊利诺斯州”作为限定词，以避免可能出现误解。当用户在地图上进行放大，进而关注相对较小的地区时，地方词库非常有用，因为此时文章本质上更着重于地方特点。在这种情况下，有关提供者的知识在克服地理 / 地理模糊性时极为有用。

还可以把地方词库看成地名分辨使用的“辨别背景”。在地名分辨使用的相关流行策略 [2,27,31,45] 中，辨别背景被放在某个层次的地理本体范围内，此时需要找到可分辨文档内众多地名的地理区域。例如，Web-a-Where[2] 通过文档中多种形式的层次证据实现此类方法。证据包括最小辨别背景和邻近地名的包含关系（比如“马里兰州帕克分校”）。使用纳入地名词典内的层级结构以及各地名的置信度得分的简单评分算法后，系统找到了文档的地理焦点。Ding 等人 [6] 使用了类似的方法。MetaCarta[33] 和谷歌图书搜索没有使用计算地理焦点的概念，因此需要用户自己确定焦点。除了使用内容的位置外，Mehler 等人 [28] 还把文档与提供者的地点关联起来，有时候这等同于使用日期栏。注意，找到最小辨别背景背后的中心假设是，拟分析的文档只有一个地理焦点，这个地点在分辨该焦点内的地名时有用，但在分辨附带提及的较远的地名时没用。

还请注意，地方词库只是报亭使用的诸多地名分辨技术之一。由于事实上某些特性与多个纪录关联，即 $|L(f)| > 1$，所以需要该词库。具体来说，报亭通过启发式的过滤器分辨此类模糊的参照。这些过滤器会为依照人类阅读文章的方式为每个参照选择一组最可能的匹配。这些过滤器依赖于报亭最初的假设，即文章中的地点在地理距离、文章距离 [19] 和层次包含关系方面为彼此提供证据。“对象容器过滤器”便是其中的一种过滤器。通过指明包含关系的关键词或标点符号（比如“f_1 in f_2” 或者“f_1, f_2”），该过滤器找出了文章中被隔开的地理特征对 $f_1, f_2 \in EFV$。如果它发现了有地点对 (l_1, l_2) 符合 $l_1 \in L(f_1)$, $l_2 \in L(f_2)$ 且 l_1 在 l_2 之内，那么 f_1 和 f_2 被分别消歧为 l_1 和 l_2。举例来说，设 f_1 =“Brooklyn”，f_2“NYC.”。同时，设 $L(f_1)$ = {“Brooklyn, New York City,” “Brooklyn, Shelby County’’}，$L(f_2)$ = {“New York City, New York County,” “North Yorkshire County, U.K.”}。我们现在可以对 f_1 和 f_2 进行消歧，分别得到 l_1 =“Brooklyn, New York City”和 l_2 =“New York City, New York County.”。这种消歧得到了报亭观察结果的证明。在观察结果中，文章中出现的位置靠近、地理上邻近和层级关系明显的特征对不大可能是偶然发生的。在该策略的另一个例子中，当查询涉及多个地点列表时，报亭设法使用相近性、同级关系和显著性线索消歧。[1,21]

评价。为了获取报亭的地理标记的性能，可以通过把“层”参数设置为“地点”而不是“图标”使报亭展现地点的实际名称，而不是地点上的新闻类别图标。采用这种方法后，便可检测错误的地理 / 地理解释（比如把“洛杉矶”放在“智利”，而不是“加利福利亚”）以及把非地理名称归类为地理名称的错误（比如“南非”的“乔治”，而不是 2012 年在“佛罗里达州奥兰多市”发生的凯西·安东尼涉嫌杀女案审判中的“乔治·安东尼”），但反之不成立。

不仅如此，把鼠标放在地点 l 的名称 n 上时（在“地点”和“图标”层中），报亭会生成一个迷你地图，并在地图和迷你地图中用蓝球标记标出具有相同名称 n 的所有其他地点 k，使得至少有一个文章簇会与 k 相关联。这张迷你地图可以让报亭很快地发现地理标记错误。研究人员现在正在研究如何使用这种信息来学习得出更好的分类器。对于特定地名 n 的任何解释 k，只要一篇文章与解释 k 相关联，即使 k 可能不正确，系统也会认为 k 是 n 的解释，这样就把决定权放在了用户手里。基于 n 的任何解释 k 来标出系统认为提及特定地点 n 的所有文章后，蓝球可让报亭避免可能出现的地名分辨错误。根据报亭的一项假设，即至少有一篇文章与解释相关联（假设较低精度的地名识别会有 100% 的召回率），我们检查了所有提及 n 的文章以得出正确的解释。结果我们发现，对于在地名词典中列出的某地点，如果地点的解释精度较低但无任何遗漏，那么报亭在地名分辨时会达到 100% 的召回率。注意，从某种角度来说，报亭也正在对其响应进行排名，其中排名最高的响应与主地图上所查询的地点相关联，而排名较低的响应则与迷你地图相关联。

Lieberman 和 Samet 使用了人工制作的文章语料库进行了实验。他们的实验结果 [18] 说明，报亭的地名识别 [18] 和地名分辨 [19] 流程优于路透社的 OpenCalais 和雅虎的 Placemaker。它们均为不公开源码的商业产品，提供了公开的网络 API 支持自动对文档进行地理标记。同时，MetaCarta 系统 [33] 提供了相似的功能，可识别文本文档中的空间线索词（比如“city of"）、特定格式的邮寄地址以及用文本描述的地理坐标。

经验教训

构建报亭的经历教导我们，地名识别和辨别中的地理标记任务比我们最初预想的要复杂得多。例如，报亭的地理标记器本来可以使用更多

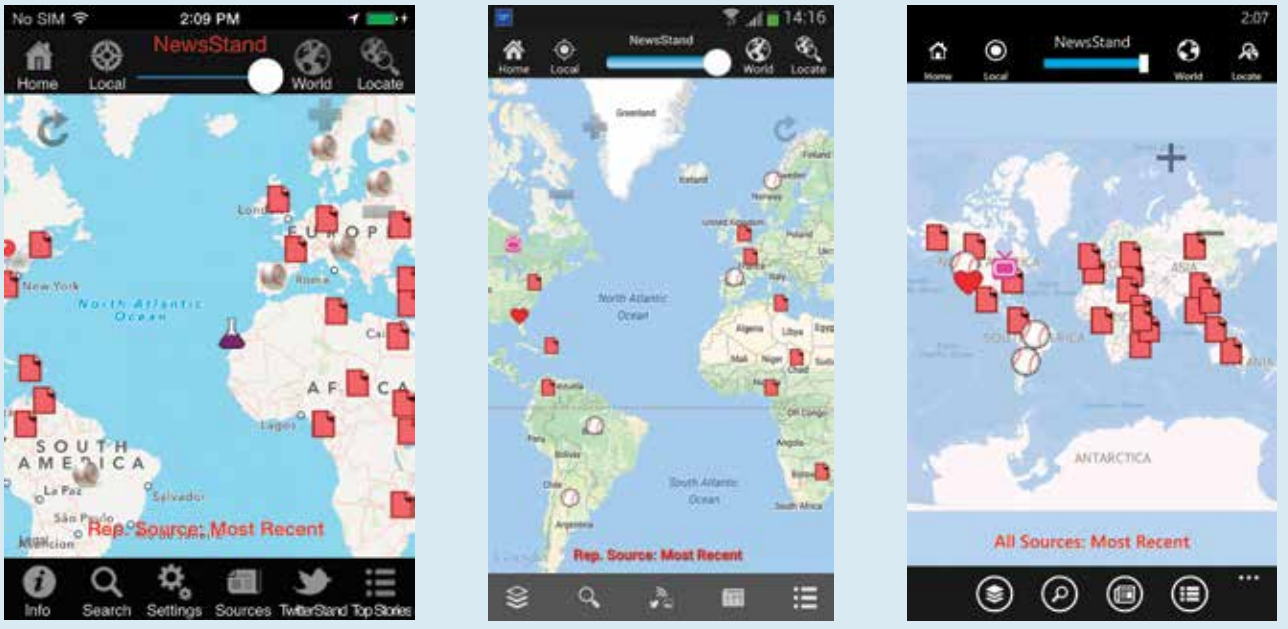
文档含有的语义提示来提高地理标记的正确性（比如地标和河流）。不仅如此，对TF-IDF框架进行修改，把各个空间上同义的搜索项合为一个搜索项而不是当成不同的搜索项后，也可以使用地理信息改进新闻文章的聚类。主要的难点在于评价报亭在上述任务中的性能。比较报亭和其他系统意味着不得不使用名为“语料库”的标准数据集。我们对地理标记任务的两个模块均进行了对比。我们把重点放在召回率而不是准确率上，最后得到了极好的结果。[18,19] 然而，这种评价方法有两个缺陷：数据集太小；而且由于新闻和语言一直变化，“语料库就如同僵化的库”。新闻数据的特征是流动的数据。评价更应该通过采样的方式进行，就如同检验 / 质量控制工作中那样。我们打算在未来这么做。

在网页浏览器中，报亭很好地使用了谷歌地图提供的地图 API 来展现主题。我们还对它进行了修改，使之可以适配必应地图和谷歌地球插件。尽管由于支撑平台的数量有限，插件导致了一些显示问题。报亭还被移植到带有支持手势的触摸屏界面的设备上（比如智能机和平板电脑）。虽然其中的用户界面有所不同，但用户可通过其中的网页浏览器进行使用 [42]。不仅如此，我们还开发了 [38] 在 iPhone、Android（安卓）和 Windows Phone 平台上使用的应用（见图 6）。报亭没有“公开”的 API，不过它的很多功能以及适配不同智能机平台的能力均使用了它的“私有”API。

由于基于浏览器的网络环境和移动设备的原生应用环境存在差异，所以需要改变用户行为或习惯。例如，网页上以地图为中心的应用最好只有一个页面，这意味着外部链接（比如通往报亭的新闻文章的链接）最好在单独的浏览器页签中打开，以保存报亭应用及其状态。如果新闻文章在相同的页签中打开，就无法实现这一功能。在单独的浏览器页签中打开外部链接也会导致不想要的结果。一个具体的例子是，用户不能使用“后退”键返回应用和应用之前的状态。与此相反，他们必须显式地关闭新打开的页签。此时，调用页签及其状态隐式地保存了。在原生的应用环境，此类问题不会发生。原生应用可以协调多个窗口之间的转换，从而提供更加友好的用户互动。不过，也牺牲了某些能力。比如在我们的例子中，一次只能打开一个链向新闻文章的外部链接。

在把报亭移植到多种不同的移动 / 智能机平台时，我们发现底层地图 API 的实现并没有遵守经典的制图原则。结果是，进行某些操作时（比如缩放和平移）会出现一致性问题。例如，一旦地点的名称出现在地图上后，在用户继续放大或平移时，只要该地点仍在窗口内，名称就应该能继续显示。[40] 有趣的是，移动和智能机平台上的某些地图应用不支持缩小后在整个屏幕上查看整个世界（比如谷歌地图和苹果地图的移动 / 智能机地图 API），因此即便整个世界在“当时”的地图 API 中已经存在，也需要继续平移才能看到世界的其他部分。[40] 在报亭中，这种现象尤其让读者厌烦，因为读者希望看到整个世界正在发生的事件。[40] 迷你地图部分缓解了这一问题，对于使用标题信息提示

图 6.报亭应用的截图：(a) iPhone、(b) Android（安装）、(c) Windows Phone 平台



框突出显示的特定文章，它用橘球标出了其中提到的所有其他地名。

在设计用户界面时，我们不得不考虑使用手势界面后无法在设备上悬停，这意味着在支持手势的平台上将不得不改变一些功能的实现方式。具体来说，当定点设备经过一个位置时，悬停支持用户观察当时正在显示或展开的现象的空间变异。手势界面要求使用轻拍（tap）或点击（click）来触发这种显示行为，因为手指在某个区域内的连续运动会被解释为一次轻拍或点击。因此，很难观察到空间变异。另一方面，缺少悬停意味着从地图位置 *l* 到另一地图位置 *b* 的过渡可以通过轻拍 *b* 处实现。相比之下，使用悬停进行从 *l* 到 *b* 的过渡时，可能需要采取某些特定的行为，而且这些行为会破坏系统的当前状态。

进行快速地图注记时，面临的设计挑战是把迷你地图放在靠近标题提示框和相关信息框的位置处。在注记的动态展现中，也会面临这种挑战（比如图 3 中的疾病名称，关键词以及人名和品牌名）。我们的目标是在平移和缩放时，按照互动的速度进行快速地图注记。这点可以通过为动态地图注记 [30] 开发且纳入 PhotoStand 系统的技术予以实现。[37]

结论

我们回顾了报亭系统的设计目标和功能。报亭系统使用地图来阅读网络中的新闻，并发挥了空间同义词的作用。报亭证明，从新闻文章中抽取地理内容会发掘之前未曾见过的信息维度，新闻确实可以被当成东南西北的首字母缩略词【NEWS（North，East，West，South）】。在互联网上，有地理标记的内容越来越流行，这使得在其他知识领域的系统中出现了与报亭相似的，令人叹服的应用。例如，情感 / 内容分析可以解释不同国家或不同语言的人群对同一篇新闻报道的不同理解方法，也可根据新闻、推文或其他数据摘要源进行热点分析。不仅如此，报亭还为新兴的计算新闻学领域做出了贡献。[8]

报亭现在为10,000个新闻源构建了索引，每天处理约50,000篇新闻文章。

未来的研究包括使用地图查询界面通过代表性的图片（比如 PhotoStand[37]）、视频或音频剪辑访问其他媒体。我们也正在设法纳入其他新闻源和信息源。例如，我们已经在报亭中纳入了 Twitter 的推文，由此创建了 TwitterStand 系统 [44]。该系统的理念是发掘大量的新闻文章作为某种类型的聚类用语料库，以便长度非常短、信息非常稀疏的推文使用已有的新闻簇进行聚类。在这一方法中，有趣的一面是由于推文长度相当短，它们往往很少有或没有地理内容。但是，当它们被聚类后，它们继承了地理信息，而这些地理信息则与推文关联的簇的地理焦点存在关联关系。我们从中发现的新结果是，焦点现在是用户的推文中涉及的地理区域，而不是用户发出推文时所在的地理区域（当发出推文的设备具有 GPS 功能时，很容易发现用户的所在位置）。在编写有关未来事件的推文时，这种焦点相当有用 [14]，但必须谨慎选择关注谁的推文。[11]

鸣谢

本文基于 Teitler 等人之前的论文。[46] 本文的研究资金部分来源于美国国家科学基金会（项目号 IIS-07-13501、IIS-08-12377、CCF-08-30618、IIS-10-18475、IIS-12-19023 和 IIS-13-20791）、美国住房和城市发展部政策发展和研究办公室、微软研究院、谷歌研究院、英伟达（Nvidia）、爱尔兰科学基金会 E.T.S. Walton 访问学者奖（E.T.S. Walton Visitor Award of the Science Foundation of Ireland）以及梅努斯镇爱尔兰国立大学地理计算国家中心。我们还要感谢 Larry Brandt、Jim Gray、Keith Marzullo、和 Maria Zemankova 的支持。

参考资料

1. Adelfio, M.D. and Samet, H. Structured toponym resolution using combined hierarchical place categories.In *Proceedings of the Seventh ACM SIGSPATIAL Workshop on Geographic Information Retrieval* (Orlando, FL, Nov. 5).ACM Press, New York, 2013, 49–56.
2. Amitay, E., Har'El, N., Sivan, R., and Soffer, A. Web-a-Where:Geotagging Web content.In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, U.K., July 25–29).ACM Press, New York, 2004, 273–280.
3. Aref, W.G. and Samet, H. Efficient processing of window queries in the pyramid data structure.In *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (Nashville, TN, Apr. 2–4).ACM Press, New York, 1990, 265–272.
4. Baldwin, B. and Carpenter, B. *Lingpipe*; http://alias-i.com/lingpipe/
5. Chum, O., Philbin, J., Isard, M., and Zisserman, A. Scalable near-identical image and shot detection.In *Proceedings of the Sixth ACM International Conference on Image and Video Retrieval* (Amsterdam, The Netherlands, July 9–11).ACM Press, New York, 2007, 549–556.
6. Ding, J., Gravano, L., and Shivakumar, N. Computing geographical scopes of Web resources.In *Proceedings of the 26th International Conference on Very Large Data Bases* (Cairo, Egypt, Sept. 10–14).Morgan Kaufmann, San Francisco, 2000, 545–556.
7. Duda, R.O., Hart, P.E., and Stork, D.G.*Pattern Classification, Second Edition.* Wiley Interscience, New York, 2000.
8. Essa, I. *Computation + Journalism:A study of Computation and Journalism and How They Impact Each Other*; http://www.computation-and-journalism.com/
9. Francis, W.N.A standard corpus of edited present-day American English.*College English 26*, 4 (Jan. 1965), 267–273.
10. Freifeld, C.C., Mandl, K.D., Reis, B.Y., and Brownstein, J.S.HealthMap:Global infectious disease monitoring through automated classification and visualization of Internet media reports.*Journal of the American Medical Informatics Association 15*, 2 (Mar. 2008), 150–157.
11. Gramsky, N. and Samet, H. Seeder finder:Identifying additional needles in the Twitter haystack. In *Proceedings of the Fifth ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (Orlando, FL, Nov. 5).ACM Press, New York, 2013, 44–53.
12. Hjaltason, G.R. and Samet, H. Speeding up construction of PMR quadtree-based spatial indexes.*Very Large Data Bases Journal 11*, 2 (Oct. 2002), 109–137.
13. Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. Robust disambiguation of named entities in text.In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Edinburgh, Scotland, July 27–31).Association for Computational Linguistics, Stroudsburg, PA, 2011, 782–792.
14. Jackoway, A., Samet, H., and Sankaranarayanan, J. Identification of live news events using Twitter. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on Location-Based Social Networks* (Chicago, Nov. 1).ACM Press, New York, 2011, 25–32.
15. Lafferty, J.D., McCallum, A., and Peireira, F.C.N.Conditional random fields:Probabilistic models for segmenting and labeling sequence data.In *Proceedings of the 18th International Conference on Machine Learning* (Williamstown, MA, June 28–July 1).Morgan Kaufman, San Francisco, 2001, 282–289.
16. Lan, R., Lieberman, M.D., and Samet, H. The picture of health:Map-based, collaborative spatio-temporal disease tracking.In *Proceedings of the First ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health* (Redondo Beach, CA, Nov. 6).ACM Press, New York, 2012, 27–35.
17. Leidner, J.L.*Toponym Resolution in Text:Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland, U.K., Oct. 2006; https://www.era.lib.ed.ac.uk/bitstream/1842/1849/1/leidner-2007-phd.pdf
18. Lieberman, M.D. and Samet, H. Multifaceted toponym recognition for streaming news.In *Proceedings of the 34th International Conference on Research and Development in Information Retrieval* (Beijing, July 24–28).ACM Press, New York, 2011, 843–852.
19. Lieberman, M.D. and Samet, H. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th International Conference on Research and Development in Information Retrieval* (Portland, OR, Aug. 12–16).ACM Press, New York, 2012, 731–740.
20. Lieberman, M.D. and Samet, H. Supporting rapid processing and interactive map-based exploration of streaming news.In *Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Redondo Beach, CA, Nov. 7–9).ACM Press, New York, 2012, 179–188.
21. Lieberman, M.D., Samet, H., and Sankaranarayanan, J. Geotagging:Using proximity, sibling, and prominence clues to understand comma groups.In *Proceedings of the Sixth Workshop on Geographic Information Retrieval* (Zürich, Switzerland, Feb. 18–19).ACM Press, New York, 2010.
22. Lieberman, M.D., Samet, H., and Sankaranarayanan, J. Geotagging with local lexicons to build indexes for textually specified spatial data.In *Proceedings of the 26th IEEE International Conference on Data Engineering* (Long Beach, CA, Mar. 1–6).IEEE Press, 2010, 201–212.
23. Lieberman, M.D., Samet, H., Sankaranarayanan, J., and Sperling, J. STEWARD:Architecture of a spatio-textual search engine.In *Proceedings of 15th ACM International Symposium on Advances in Geographic Information Systems* (Seattle, Nov. 7–9).ACM Press, New York, 2007, 186–193.
24. Lieberman, M.D., Sankaranarayanan, J., Samet, H., and Sperling, J. Augmenting spatio-textual search with an infectious disease ontology.In *Proceedings of the Workshop on Information Integration Methods, Architectures, and Systems* (Cancun, Mexico, Apr. 11–12).IEEE Computer Society, 2008, 266–269.
25. Lowe, D.G.Object recognition from local scale-invariant features.In *Proceedings of the Seventh International Conference on Computer Vision* (Corfu, Greece, Sept. 20–25).IEEE Computer Society, 1999, 1150–1157.
26. Markowetz, A., Brinkhoff, T., and Seeger, B. Exploiting the Internet as a geospatial database.In *Proceedings on the Workshop on Next Generation Geospatial Information* (Cambridge, MA, Oct. 19–21, 2003).
27. Martins, B., Manguinhas, H., Borbinha, J., and Siabato, W. A geo-temporal information extraction service for processing descriptive metadata in digital libraries. *e-Perimetron 4*, 1 (2009), 25–37.
28. Mehler, A., Bao, Y., Li, X., Wang, Y., and Skiena, S. Spatial analysis of news sources.*IEEE Transactions on Visualization and Computer Graphics 12*, 5 (Sept.–Oct. 2006), 765–772.
29. Milne, D. and Witten, I.H.Learning to link with Wikipedia.In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, CA, Oct. 26–30).ACM Press, New York, 2008, 509–518.
30. Nutanong, S., Adelfio, M.D., and Samet, H. Multiresolution select-distinct queries on large geographic point sets.In *Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Redondo Beach, CA, Nov. 7–9).ACM Press, New York, 2012, 159–168.
31. Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., and Yang, B. The design and implementation of SPIRIT:A spatially aware search engine for information retrieval on the Internet.*International Journal of Geographical Information Systems 21*, 7 (2007), 717–745.
32. Quercini, G., Samet, H., Sankaranarayanan, J., and Lieberman, M.D. Determining the spatial reader scopes of news sources using local lexicons.In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (San Jose, CA Nov. 3–5).ACM Press, New York, 2010, 43–52.
33. Rauch, E., Bukatin, M., and Baker, K. A confidence-based framework for disambiguating geographic terms.In *Proceedings of the HLT-NAACL Workshop on Analysis of Geographic References* (Edmonton, Canada).Association for Computational Linguistics, Stroudsburg, PA, 2003, 50–54.
34. Rizzo, G. and Troncy, R. NERD:A framework for unifying named entity recognition and disambiguation extraction tools.In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (Avignon, France, Apr. 23–27).Association for Computational Linguistics, Stroudsburg, PA, 2012, 73–76.
35. Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval.*Information Processing & Management 24*, 5 (1988), 513–523.
36. Samet, H. *Foundations of Multidimensional and Metric Data Structures.*Morgan Kaufmann, San Francisco, 2006.
37. Samet, H., Adelfio, M.D., Fruin, B.C., Lieberman, M.D., and Sankaranarayanan, J. PhotoStand:A map query interface for a database of news photos.*Proceedings of the VLDB Endowment 6*, 12 (Aug. 2013), 1350–1353.
38. Samet, H., Adelfio, M.D., Fruin, B.C., Lieberman, M.D., and Teitler, B.E.Porting a Web-based mapping application to a smartphone app. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Chicago, Nov. 2–4).ACM Press, New York, 2011, 525–528.
39. Samet, H., Alborzi, H., Brabec, F., Esperança, C., Hjaltason, G.R., Morgan, F., and Tanin, E. Use of the SAND spatial browser for digital government applications.*Commun.ACM 46*, 1 (Jan. 2003), 63–66.
40. Samet, H., Fruin, B.C., and Nutanong, S. Duking it out at the smartphone mobile app mapping API corral:Apple, Google, and the competition.In *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems* (Redondo Beach, CA, Nov. 6).ACM Press, New York, 2012, 41–48.
41. Samet, H., Rosenfeld, A., Shaffer, C.A., and Webber, R.E.A geographic information system using quadtrees. *Pattern Recognition 17*, 6 (Nov./Dec. 1984), 647–656.
42. Samet, H., Teitler, B.E., Adelfio, M.D., and Lieberman, M.D. Adapting a map query interface for a gesturing touchscreen interface.In *Proceedings of the 20th International Word Wide Web Conference* (Hyderabad, India, Mar. 28–Apr. 1).ACM Press, New York, 2011, 257–260.
43. Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M.D., and Sperling, J. TwitterStand:News in tweets.In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Seattle, Nov. 4–6).ACM Press, New York, 2009, 42–51.
44. Sarma, A.D., Lee, H., Gonzales, H., Madhavan, J., and Halevy, A. Efficient spatial sampling of large geographical tables.In *Proceedings of the ACM SIGMOD Conference* (Scottsdale, AZ, May 20–24). ACM Press, New York, 2012, 193–204.
45. Stokes, N., Li, Y., Moffat, A., and Rong, J. An empirical study of the effects of NLP components on geographic IR performance.*International Journal of Geographical Information Systems 22*, 3 (Mar. 2008), 247–264.
46. Teitler, B., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. NewsStand:A new view on news.In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Irvine, CA, Nov. 5–7).ACM Press, New York, 2008, 144–153.
47. Zhou, G. and Su, J. Named entity recognition using an HMM-based chunk tagger.In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, PA, July 6–12).Association for Computational Linguistics, Stroudsburg, PA, 2002, 473–480.

Hanan Samet（hjs@cs.umd.edu）是马里兰州马里兰大学帕克分校计算机科学系、自动化研究中心以及高级计算机研究所的杰出教授。

Jagan Sankaranarayanan（sjagan@gmail.com）是加利福尼亚州库比蒂诺NEC实验室的研究员；他在本文中的研究是他在马里兰大学帕克分校高级计算机研究所任助理研究员时进行的。

Michael D. Lieberman（mike.d.lieberman@gmail.com）是马里兰州劳雷尔（Laurel）市约翰霍普金斯大学应用物理实验室的研究员；他在本文中的研究是他在马里兰大学帕克分校攻读计算机科学博士时的一部分成果。

Marco D. Adelfio（marco@cs.umd.edu）是马里兰大学帕克分校计算机科学的在读博士。

Brendan C. Fruin（bcfruin@gmail.com）是华盛顿州西雅图Zillow公司的软件工程师；他在本文中的研究是他在马里兰大学帕克分校攻读计算机科学硕士时的一部分成果。

Jack M. Lotkowski（JackLotkowski@gmail.com）是马里兰大学帕克分校的本科生。

Daniele Panozzo（daniele.panozzo@gmail.com）是瑞士苏黎世瑞士联邦理工学院（ETH）的高级研究员；他在本文中的研究是他在马里兰大学帕克分校高级计算机研究所当进修生时进行的。

Jon Sperling（jonxsperling@gmail.com）是华盛顿特区美国住房和城市发展部政策发展和研究办公室的高级研究员。

Benjamin E. Teitler（bteitler@cs.umd.edu）在本文中的研究是他在马里兰大学帕克分校攻读计算机科学硕士时的一部分成果。

译文责任编辑：崔斌