arXiv:2305.00546v1 [cs.IR] 30 Apr 2023

使网页中的变化可被发现：面向网页档案的变更文本搜索界面

LESLEY FREW,老道明大学，美国MICHAEL L.
NELSON,老道明大学，美国MICHELE C.
WEIGLE,老道明大学，美国

网页会随时间变化，而网页档案保存着网页的历史版本副本。网页档案的用户（例如记者）希望查找并查看网页随时间发生的变化。然而，目前的网页档案搜索界面并不支持这一任务。对于提供全文搜索功能的网页档案，符合搜索查询的同一网页的多个版本会分别显示，却不列出其中的变化；或者这些版本会被归为一组，导致变化被隐藏。我们介绍了一种变更文本搜索引擎，使用户能够查找网页中的变化。我们介绍了该搜索引擎后端和前端的实现，包括一种工具，它能够以动画形式在上下文中展示两个网页版本之间的变化。我们使用2016年至2020年间发生变化的美国联邦环境网页对该搜索引擎进行了评估。变更文本搜索结果页面能够清楚地显示网页中新增或删除词语和短语的时间。还可以查询倒排索引，以识别语料库中显著且频繁被删除的词项。

CCS 概念：• **信息系统 → 搜索界面**；**数字图书馆与档案**。

附加关键词和短语：网页档案、信息检索、政府文件、版本化文档集合

1 引言

我们当前与万维网互动的方式，包括使用浏览器和热门搜索引擎，往往使我们只能看到网页的最新版本。一些网页会频繁更新其内容，而另一些则相对静态。网页档案保存了这些网页的历史版本。不同的网页档案拥有不同的馆藏，而汇总这些馆藏可以增加可供查看的网页版本数量。

记者经常将网页档案作为证据。[1] 这些记者引用网页的存档副本，不仅是在网页无法访问时，也是在网页随着时间推移发生变化时 [10]。正如 Teevan[36], 所指出的，仅仅保存网页的过去版本，而不提供适当的发现方式，并不足以帮助人们找到他们正在寻找的历史内容变化。

全文搜索已被认为是网页档案中一项需求非常高的功能。 [32] 然而，如图 1 所示，许多网页档案仅提供馆藏的统一资源标识符查找，而不提供全文搜索。互联网档案馆的时光回溯机确实提供了一个搜索引擎，但它基于元数据，例如网页标题和域名，而不是全文索引。一些网页档案和单独的网络档案馆藏确实拥有全文搜索引擎。图 2 展示了其中一些搜索引擎的搜索引擎结果页，图 3 展示了这些搜索引擎当前向用户提供的查询字段。现有界面均不允许用户查询网页变化。

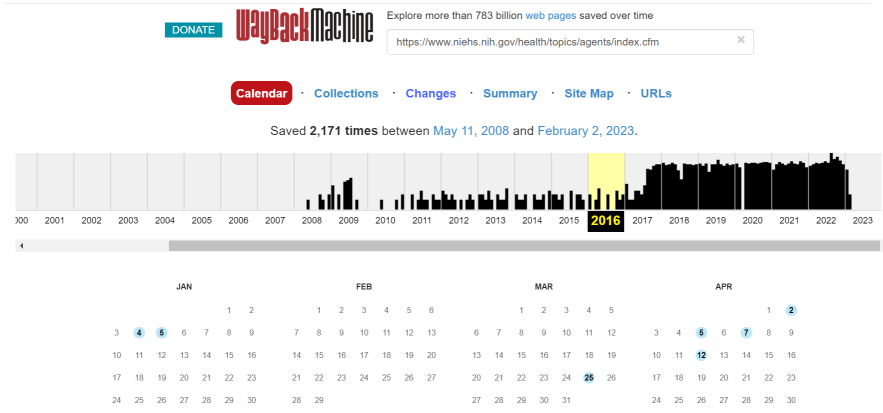
例如，2016年至2020年间，nih.gov 上的“环境因子”网页中删除了许多“污染”一词的实例，但目前的网页存档搜索界面都无法帮助用户轻松找到具体变更发生的准确日期和时间。图 2a 展示了 Archive-It 的美国国立卫生研究院网页档案馆藏中，同一网页的多个版本作为独立搜索结果列出的情况。[2] 无法明确看出为何某个结果的排名高于另一个结果，而且当前界面也没有说明这些版本之间有何差异。图 2b 展示了

[1]https://archive.org/about/news-stories/search?mentions-search=时光回溯+机器 [2]https://archive-it.org/collections/1170

葡萄牙网络档案中一个匹配的网页版本，[3] 但该网络档案实际上还保存着此网页的其他匹配版本，这些版本并未出现在搜索结果中。图 2c 展示了在互联网档案馆的 WaybackMachine 上搜索 2016年任期结束档案的结果，[4] 然而，搜索结果所链接的版本日期为 2017年1月24日，即就职日之后。稍后，我们将展示一个更有效的网页变更搜索结果页面，说明删除发生在 2月至2017年3月之间。

在这项工作中，我们为网络档案馆藏提出了一种时间型搜索引擎架构和界面，让用户能够搜索网页上的变更，并在上下文中查看这些变更。该搜索引擎架构的主要贡献是计算网页存档快照之间的变更词项。搜索引擎的主要贡献

[3]https://arquivo.pt/ [4]
https://web.archive.org/EndOfTerm2016WebCrawls/search/



(a) 互联网档案馆 Wayback Machine 统一资源标识符查找



(b) 葡萄牙网络档案馆 URI 查询

图 1. 两个示例 https://www.niehs.nih.gov/health/topics/agents/index.cfm 的网络档案 URI 查找示例。URI 查找不做 t 允许用户 t 以页面文本搜索。

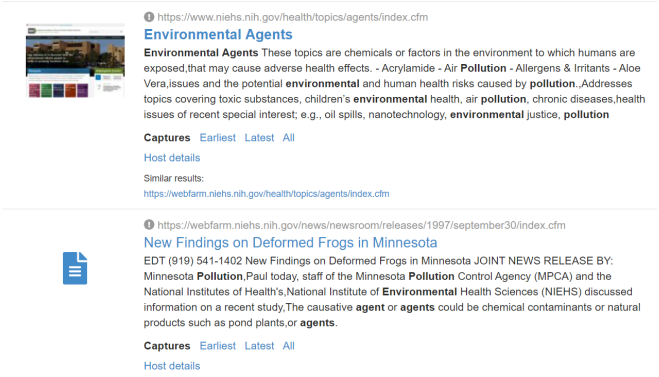
使网页中的变化可被发现



(a) 与图 3a 中查询匹配的 Archive-It 档案集合搜索引擎结果页，显示标题、统一资源标识符、日期、文本摘要、元数据、重放链接以及指向其他捕获版本的链接。图中单独展示了两个与查询匹配的捕获版本。



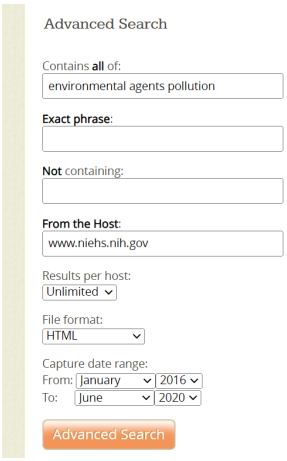
(b) 与图 3b 中的查询匹配的 Arquivo.pt 搜索引擎结果页，显示统一资源标识符、标题、日期、文本摘要和重放链接。



(c) 图 3c 中与查询匹配的时光回溯机集合搜索引擎结果页，显示统一资源标识符、标题、文本摘要、屏幕截图、重放链接以及指向更多网页存档捕获的链接。

图 2。这三个网页存档搜索界面都没有将版本分组，以展示网页随时间的变化。

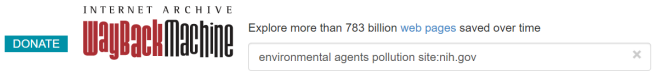
使网页中的变化可被发现



(a) Archive-It 中美国国立卫生研究院网页档案集合 1170 的高级搜索框支持全文搜索，并可按日期和域名筛选。



(b) 葡萄牙网络档案中的搜索框支持全文搜索和按日期筛选。



(c) 互联网档案馆时光回溯机中用于 2016年任期结束档案集合的搜索框支持全文搜索和按网站筛选，但一次只能搜索一个任期结束档案集合。

图 3. 一些网页档案支持全文搜索，并提供多种筛选功能，可用于规模较小的档案集合或整个网页档案。所有搜索界面都不支持搜索已从网页中删除的词语。

界面包括：可查询变化的搜索引擎；以有意义的方式将同一网页的多个版本作为一条结果显示的搜索引擎结果页；允许用户跨多个时间段查看网页文本变化的滑动差异查看器；以及允许用户在上下文中查看变化的差异动画。

变更文本搜索引擎后端使用环境数据与治理倡议（EDGI）的联邦环境网页数据集 [26] 进行了评估。用户可以在搜索引擎结果页的摘要中查看网页变化。索引流程能够识别原始 EDGI 研究中已识别词语之外的其他词语，并且与原研究的四年时间窗口相比，可以提高删除内容的时间粒度。

2 背景与相关工作

Memento协议 [37] 是网页档案的标准 HTTP 内容协商协议。用户通过实时网络上特定网页的统一资源标识符（也称为 URI-R）请求该网页，并提供一个日期时间。符合Memento协议的网络档案服务器会查询并返回最接近所请求日期时间的网页存档版本。每个网页存档版本都可以通过其独立的统一资源标识符（也称为 URI-M）直接获取。网页的存档版本也称为网页存档快照。此外，还可以查询特定服务器上某个网页的所有存档版本列表。该列表称为*TimeMap*，也称为 URI-T。网页存档快照通常被视为历史网络的一部分，而仍可使用的 URI-R 则属于实时网络。网络档案有多种，包括国家网络档案（如葡萄牙网络档案）、面向组织的订阅制网络档案（如 Archive-It），以及覆盖范围更广的网络档案（如互联网档案馆的时光回溯机）。虽然Memento协议已成为网页档案中的标准，但全文搜索的支持程度并无统一标准。网页档案包含网页的多个版本

随着时间推移，但现有的网页档案全文时序搜索引擎并未有效解决如何呈现同一网页多个版本的问题。

网页档案蕴藏着巨大的潜力，可用于分析随时间发生的变化。网页的变化速度各不相同，取决于域名和 URL 深度等特征 [3]。学术研究中常用的现有网页抓取数据集，例如 ClueWeb [28] 和 Common Crawl，[5]，旨在收集大量不同 URL 的快照。由于这些数据集并非旨在收集同一网页的多个版本，因此无法保证定期抓取任何特定网页，这会妨碍对随时间变化进行分析。此外，要使网页变化分析具有实际意义，这些网页上的变化必须是有意义的。显然，利用网页档案研究主题随时间发生的变化仍有局限，因为目前无法搜索发生特定变化的网页。

为了充分利用网页档案来考察随时间发生的变化，时间类搜索界面必须扩展功能，以支持搜索词语和短语的变化。这些搜索界面还需要向用户有效展示这些变化。最后，搜索工具需要能够在与现有数字人文研究相关的网页数据集中找到有意义的变化。

2.1 网页档案的时间搜索引擎

并非所有网页档案都提供全文搜索功能。一些网页档案，例如葡萄牙网络档案[25], 、英国网络档案馆[6] 以及 Archive-It 中的档案集合[7]，确实支持全文搜索。这些系统彼此独立；目前没有一种全文搜索引擎能够搜索多个网页档案中某个网页的所有版本。创建网页档案时间搜索引擎的一项主要挑战在于，每个归档网页可能有多个版本。目前还没有适用于版本化文档集合的现成信息检索系统。创建此类系统时，需要解决索引、呈现搜索结果以及在上下文中显示变更等问题。

版本化文档集合的索引方式对于发现不同版本之间的变更至关重要。Berberich 等人 [5] 开发了一种时间合并框架，按时间顺序排列文档的所有版本，并为每份文档指定一个有效期范围，而非单个日期时间。这样便可以量化不同版本之间的文档变更量。为节省空间，索引中会合并相似度达到特定阈值的文档，但搜索引擎结果页中并未合并各个文档版本，也无法按变更进行搜索。虽然 Berberich 开发有效期范围框架时尚未实现时间范围，但日期范围字段如今已成为 Apache Lucene [35], 这一首屈一指的开源搜索后端的标准组成部分。[8]

少数时间搜索引擎研究了如何在搜索结果中呈现网页的多个版本。Melo 等人 [25] 在葡萄牙网络档案搜索系统的后端添加了一个参数，用于限制搜索结果页面上显示的版本数量。Kiesel 等人 [20] 对其个人网络存档本地化搜索系统进行了定性评估，发现包含网页的所有版本会使搜索结果页面显得杂乱，从而影响可用性。Major [24] 也指出，在搜索结果中重复出现同一统一资源标识符（URI）的捕获版本是一个问题。另一方面，Jackson 等人 [15]选择在搜索结果页面中按时间顺序纳入网页的所有相关版本。这是因为对网页版本进行分组会掩盖随时间发生的变化。在所有方面。

[5]https://commoncrawl.org/ [6] https://www.webarchive.org.uk/ukwa/ [7] https://archive-it.org/ [8] https://lucene.apache.org/

在这些系统中，无法按更改内容进行搜索，而查看不同版本之间差异的唯一方式是手动检查。

2.2 显示文档版本之间的变化

网页档案的时间搜索引擎不会将网页不同版本之间的变化纳入搜索结果，但也存在一些其他工具，可帮助用户查找并查看已知网页的变化。Sherratt 等人 [34] 创建了一款工具，作为 GLAM Workbench 网页档案 Jupyter 笔记本集合的一部分，帮助用户查找特定网页版本历史中的变化。由于此工具不会为任何网页内容建立索引，因此必须逐个搜索 URL。该工具也仅提供线性搜索选项，这会降低搜索速度。搜索结果会在上下文中突出显示查询词，但不会呈现该词在网页先前版本中的上下文。另一款工具 WikiBlame，[9] 允许用户搜索特定维基百科网页中的变化。虽然此工具支持二分搜索，但它不会为任何网页内容建立索引，因此一次只能搜索一个网页，而不能搜索一组相关网页。维基百科在其版本历史查看器中提供了差异工具，[10] 让用户可以在静态上下文中查看变化。

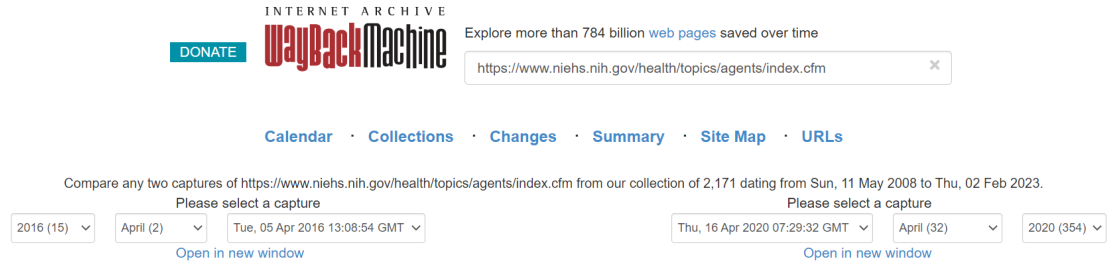
互联网档案馆 Wayback Machine 允许用户使用其 Changes 工具比较网页的两个不同版本。[11] 如图 4 所示，该工具基于环境数据与治理倡议的 Web-Monitoring-Diff 套件构建。[12] 该服务目前未与任何可供用户搜索不同版本间变更的工具相连，并且只允许用户比较来自互联网档案馆的版本。目前尚无任何版本的 Changes 工具能够让用户比较分布在多个档案馆中的网页版本。Changes 工具一次只能比较网页的两个版本，目前比较结果以静态并排的形式呈现。

除网页档案之外，其他类型的版本化文档集合也催生了支持比较多个文档版本的工具。例如，Henley 等人 [13] 开发了一款名为 Yestercode 的工具，程序员可以使用滑块在代码的不同版本之间切换，并显示相邻版本之间的差异。协作文档撰写工具 DocuViz [38]旨在帮助用户直观地了解文档的演变过程，而 Perez-Messina 等人 [29] 则开发了一款工具，用于可视化协作文档中文本片段的来源。

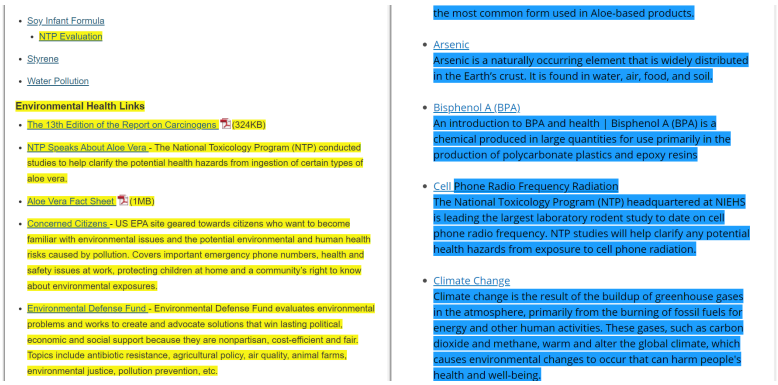
精心的设计选择对于帮助用户发现变化至关重要。前注意加工是设计中采用的一项核心技术，有助于快速检测变化，例如在估算情境中 [12]。Diffamation [7] 是一种系统，它通过动画呈现文本变化，并以并排显示作为补充。该系统通过带导航功能的并行动画，展示文档两个版本之间的所有文本变化，帮助用户了解哪些位置、哪些内容以及有多少文本发生了变化。SlideDiff [8] 是另一个利用文本动画展示文档版本差异的系统。SlideDiff 展示幻灯片演示文稿不同版本中的文本和媒体变化，甚至显示模拟鼠标光标，以吸引用户注意每处变化的位置，并通过更逼真的动画帮助用户推断编辑者的意图。这些目标同样适用于查看和理解网页上的编辑。Jatowt 等人 [17] 创建了 Journey to the Past 框架，其中包含一个浏览器，允许用户在网页档案中搜索网页的变更，并为符合查询词的差异添加动画。该浏览器的界面类似于视频回放，用户可以通过按钮在不同版本之间导航并控制动画重放。由于该框架不包含索引，搜索变更后的词语或短语必须逐页进行，而不能在整个集合层面进行。Adar 等人 [2] 开发了

[9] http://wikipedia.ramselehof.de/wikiblame.php [10] https://en.wikipedia.org/wiki/Help:Diff [11] https://web.archive.org/web/changes/ [12] https://github.com/edgi-govdata-archiving/web-monitoring-diff

使网页中的变化可被发现



(a) Wayback Machine Changes 工具要求用户知道网页两个版本的日期和时间，才能进行比较。



（b）时光回溯机 Changes 工具帮助用户查看新增内容（蓝色）和删除内容（黄色）。左侧以黄色显示的术语“污染”已被删除。

图 4。互联网档案馆的时光回溯机 Changes 工具向用户展示两个网页存档捕获之间的差异。它只能用于该网络档案中的网页存档捕获，且无法搜索这些变化来查找日期和时间。

Zoetrope 系统允许用户搜索单个网页版本中的术语，并以定格动画风格展示这些差异随时间的变化。该系统依赖本地爬取，查询依赖于文档对象模型中的最接近匹配，而不是索引；将全文搜索与 Zoetrope 的可视化功能相结合被确定为未来的工作。

2.3 联邦网站上的变化

联邦网站的归档，尤其是在总统任期结束时进行归档，是多家组织承担的一项重要工作。任期结束网页档案由五家组织合作创建，其中包括互联网档案馆和美国国会图书馆 [31, 33]。该网络档案提供全文搜索功能，[13] 但每次任期结束时的网页爬取都只包含每个网页的一个捕获版本。Phillips 等人 [30] 比较了 2008年和 2012年任期结束时的档案集合，以识别爬取日期和网页地址的变化，但没有分析具体术语。Nost 等人 [26], 代表环境数据与治理倡议（EDGI），利用互联网档案馆的馆藏，比较了 2016年与 2020年之间 56 个预先选定的环境相关术语和短语的变化。他们整理出一份约含 10,000 个网页的清单，这些网页在 2016年和 2020年都有版本。EDGI

[13]http://eot.us.archive.org/搜索/

的分析侧重于呈现档案集合层面的趋势，而非创建一个供用户浏览单个网页变化的界面。2008/2012 和 2016/2020 两项分析都依赖于对比间隔四年捕获的版本，但这些网页在多个尚未进行变化分析的网页档案中还存在许多其他版本。

3 设计动机

对于许多用户来说，建立对网页档案的心智模型并学习如何浏览历史网络并非易事。就在 2019 年，Abrams 等人 [1] 发现，用户难以分辨自己访问的是实时网络还是历史网络。他们还发现，用户对网页档案缺乏了解，对其成功使用的阻碍不亚于低效的用户界面。因此，全文搜索是一项高级功能，只有对历史网络有深入了解的用户才能从中受益。

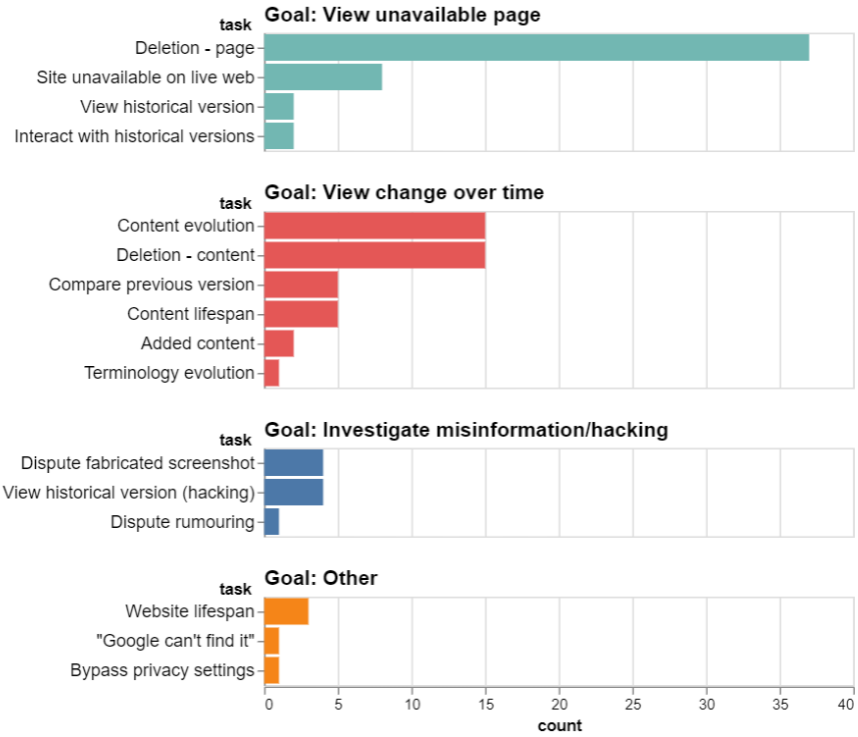


图 5. 记者使用网页档案作为文章证据时，最常见的两个目标是查看无法访问的网页，并查看网页内容 d 内容随时间的变化。

记者是网页档案全文搜索的一类潜在用户。此前尚无研究分析记者如何使用网页档案 [27]。Teevan曾开展一项研究，分析普通用户为何希望访问已消失的网页 [36]。我们采用与Teevan相似的方法开展了一项简短的自然情境研究，以归纳记者使用网页档案时的任务，并确定全文搜索是否能为他们提供帮助。

我们使用 GNews Python API 搜索 Google 新闻中包含“时光回溯机”这一词组的报道，[14] 并在2022年5月至7月期间每两周收集500篇符合该查询条件的文章。在收集到的500篇文章中，有106篇将网页档案用作证据。我们根据用户任务对这些文章进行了人工分类。记者在文章中说明了自己如何使用网页档案，我们据此创建了任务编码。与Teevan的做法类似，一位作者对数据集进行了编码。如图5所示，记者使用网页档案查看无法访问的网页，也经常用它们查看随时间发生的变化。这些用户希望查看词语和短语的新增内容、删除内容，以及相关内容的存续时间 [10]。

目前，要查找术语和短语的新增与删除，只能通过手动检查网页的多个版本来实现。记者对网页档案已有较强的心智模型，因此，若搜索引擎能让他们找到内容随时间发生的变化，他们的工作就会更轻松。

虽然我们研究的用户群体只有记者，但能够从此类工具中受益的并非只有记者。IIPC 访问工作组 [11] 指出了其他专业用户，包括研究计算与人文学科领域中随时间变化的研究人员、调查数据演变的专业人士（例如旅游业或房地产领域），以及处理民事商标案件和专利案件的律师。

与其把搜索结果中显示网页的多个版本视为一种阻碍，不如将查找并展示这些版本之间的变化作为网络档案时间搜索引擎的主要功能之一。

4 架构

如图6所示，变更文本搜索引擎的架构分为三个层级。首先，必须获取具有多个版本的网页。接下来，对网页的文本内容和重放内容建立索引。完成初始索引后，计算文本内容中的变化。最后，搜索引擎的用户界面为用户提供发现网页变化并在上下文中重放这些变化的方式。

4.1 文档获取

变更文本搜索引擎所需的文档必须采用 WARC 格式。虽然用户可以重放公众网络档案中的馆藏内容，但无法获取原始 WARC 文件。因此，如果没有能够将 URI-M 转换为 WARC 文件的工具，就无法将这些公众馆藏编入变更文本搜索引擎的索引。Hypercane [19] 是一款用于与网络档案馆藏交互的工具。Hypercane 工具的一项功能是根据给定的 URI-M合成WARC 文件。Hypercane 输出的 WARC 文件包含网页的原始 HTML，以及所有可用的嵌入资源。HTML 对创建网页内容变更的倒排索引必不可少，而嵌入资源则是进行合并回放、以原有情境展示网页版本之间差异的必要条件。拥有现成 WARC 集合的用户和组织无需执行额外的文档获取。

4.2 索引

为了支持查询和重放，必须对文档集合进行索引。生成的第一个索引是 Apache Lucene 索引，其中包含网页的分词内容以及其他网页元数据。英国网络档案 WARC 索引器 [14] 是一个现有工具，可为 Lucene 正确索引 WARC 文件。Apache Solr 提供

[14]https://news.google.com/，https://github.com/ranahaani/GNews/

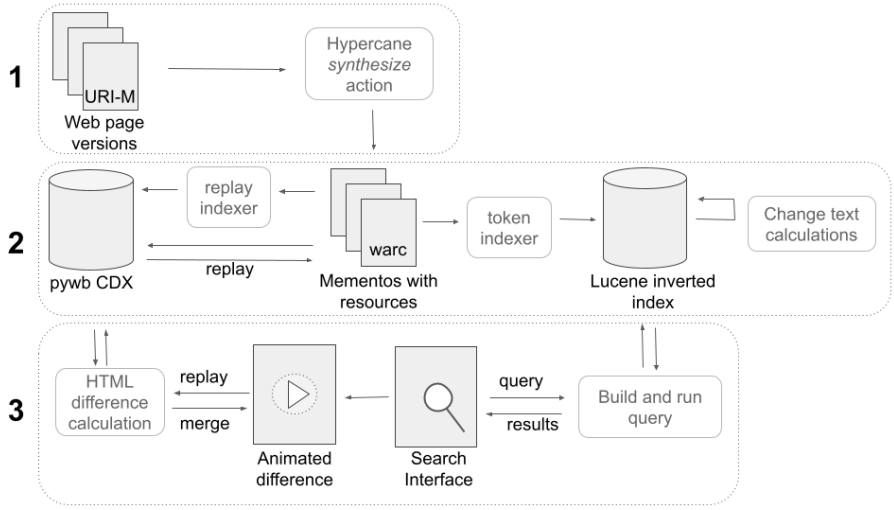


图 6. 变更文本搜索引擎的架构。第 1 层由文档获取组成。第 2 层由文档和索引组成。第 3 层由用户界面组成。

一个用于操作 Lucene 的界面。[15] 我们扩展了源自 SolrWayback [9] 的 Solr 模式，以便在索引之前支持新的变更文本字段（分别用于新增词项和删除词项集合）。我们编写了变更文本计算脚本来填充这些新字段。这段 Java 代码直接与 Lucene索引交互，计算新增词项和删除词项集合，以及 Berberich 等人 [5]定义的文档时间有效范围。文档按照其规范化 URL 分组；该字段已在初始索引过程中填充。

生成的第二个索引是 PyWB [21] CDX，其中包含重放网页所需的信息。由于 PyWB 的默认行为是将所有已编入索引的 WARC 文件复制到同一个文件夹中，因此我们关闭了此行为，以减少磁盘空间占用。只需将包含 WARC 文件的文件夹添加到 PyWB 配置文件中，即可成功重放，而无需重复存储。

4.3 用户界面

用户可以通过搜索界面与变更文本词项索引进行交互。我们使用 Solarium（一种 PHP Solr 界面）构建了该界面。用户可以查询删除词项、删除短语、新增词项或新增短语。无论词项是从网页中完全删除还是仅部分删除，都可以检测到被删除的词项和短语。变更类型以及词项或短语会被转换为针对相应变更文本字段的有效 Lucene 查询，然后返回搜索结果，并显示在搜索引擎结果页中，如图 7 所示。

<escape>[15]https://solr.apache.org/ [16]
https://github.com/solariumphp/solarium</escape>

使网页中的变化可被发现

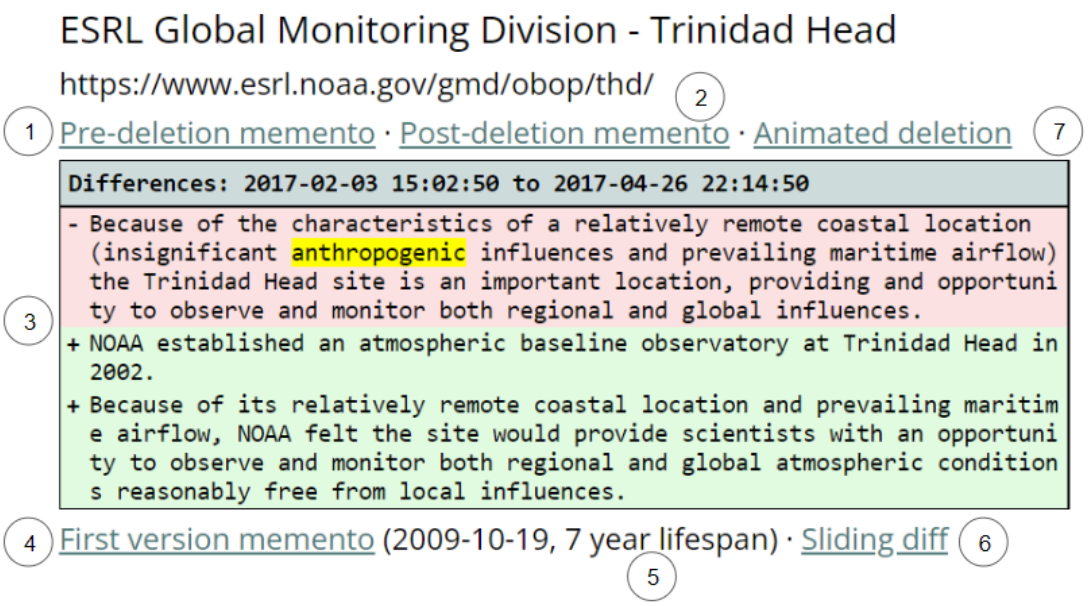


图 7. 变更文本搜索界面。1、2 和 4：指向各个网页存档快照的单独重放链接；3：删除前后版本之间的差异；5：内容生命周期计算；6：指向该网页所有已索引版本的滑动差异查看器的链接；7：指向删除动画的链接

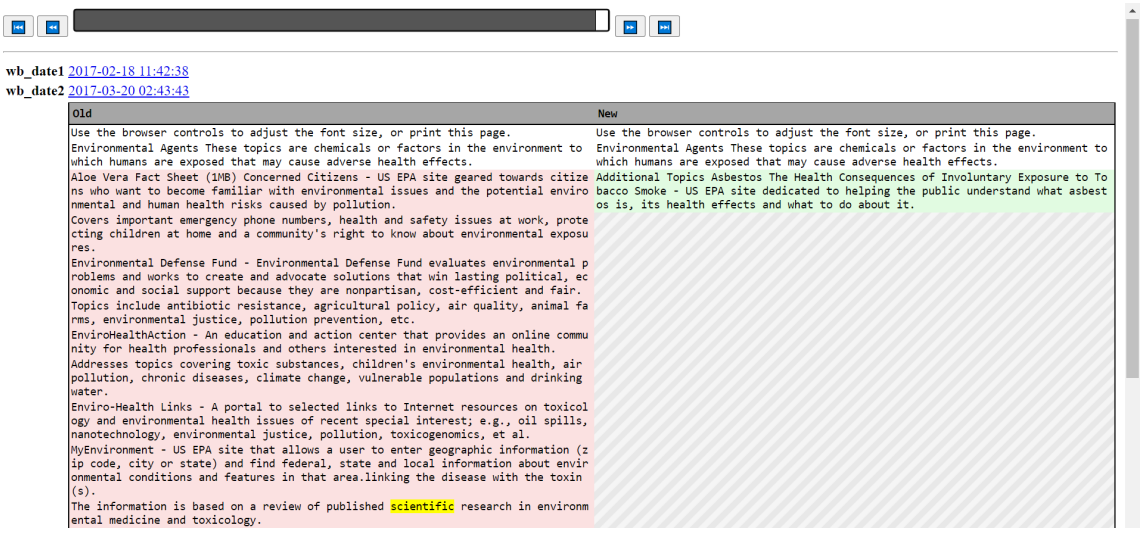


图 8. 滑动差异查看器显示，术语“科学的”于 2017 年从网页 https://www.niehs.nih.gov/health/topics/agents/index.cfm 中删除，同时展示了删除内容的上下文。

使网页中的变化可被发现



图 9. 动画显示了网页 http://www.fws.gov/ENDANGERED/permits/index.html 上短语“濒危物种”的删除。

变更文本搜索界面让用户能够了解同一网页多个版本之间的差异，这与图 2 所示的现有网络档案界面不同。每条搜索引擎结果都会显示该词语何时新增和/或删除，并附上一段摘录，其中用不同色调呈现与查询最相关的两个连续版本之间的差异。搜索引擎结果包含通过标准 PyWB 重放分别回放这些版本的链接、一个用于比较新增与删除之间所有网页版本的链接（如图 8 所示），以及一个用于同时回放删除前后版本的链接（如图 9 所示）。在第一个工具——滑动差异工具中，用户可以使用快进和倒带按钮跳过内容相同的网页存档快照。我们使用 Lucene 索引中的纯文本内容创建了这个滑动差异查看器。该查看器通过 Solarium 查询、PHP 差异库 [6], 和一些额外的 JavaScript 实现功能。第二个工具，即双重回放工具，

会以动画形式展示上下文中的差异。此工具的实际演示见 https://youtu.be/qHSVvcubuYo。动画中的差异通过不同色调呈现。动画本身以及高亮文本的颜色都有助于用户在无需刻意思考的情况下处理变化信息。与时光回溯机上的静态 Changes 工具不同，动画会依次跳转到每一处变化。动画也旨在营造变化实时发生的错觉。为制作这一动画，我们使用 EDGI 的 Python HTML 差异库计算并合并静态上下文中的差异，随后扩展代码，生成 HTML 和 JavaScript，以动画形式呈现合并后的网页。借助 PyWB 的无横幅回放，我们支持成功计算源自不同网页档案的网页版本之间的差异。为了吸引用户关注每一处变化，网页会逐一跳转到各处变化，每次为一处变化添加动画，并在跳转到下一处变化前暂停。

5 评估

5.1 查找联邦政府网页中的变化

EDGI 计算出的 2016年至2020年间发生变化的联邦政府网页，是评估变更文本搜索引擎的理想数据集。EDGI 数据集包含约 40,000 个网页地址。其中，约 10,000 个网页在互联网档案馆中有 2016年和2020年的版本。首先，我们通过考察其他网页档案，扩充数据集，以纳入更多原始网页。接着，我们筛选结果，只保留 HTTP 状态码成功（HTTP 200 OK）的网页用于索引。最后，我们优先处理已知发生术语更改的网页，并在四年时间窗口内为这些网页找出具有显著变化的其他版本，以供索引。

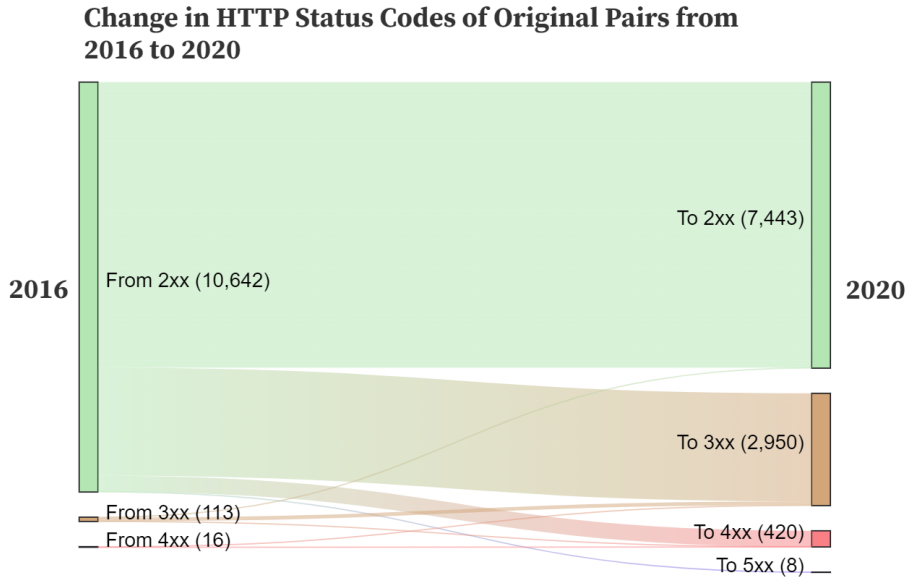
EDGI 检查了每个受监测网页，以确定其在 2016年上半年和 2020年上半年是否都存在捕获版本，并将这种情况定义为配对页面样本。在互联网档案馆中，40,000 个受监测网页里只有约 10,000 个拥有配对的网页存档快照，因此还需利用其他网页档案检查约 30,000 个网页是否存在配对的网页存档快照。我们使用 MemGator [4],这一纪念版本聚合服务，为这 30,000 个网页地址分别生成了 TimeMap。检查这些时间地图后发现，其中约 8,500 个网页确实找到了此前未发现的配对网页存档快照。有趣的是，这些新增配对中约有一半存在于互联网档案馆。这些网页存档快照可能是在某种延迟或禁运期之后添加的。其余配对则至少有一个网页存档快照来自互联网档案馆以外的网页档案，用于构成 2016/2020 配对。

在最初约 10,000 对配对的网页存档快照中，如图10a所示，约 75% 的快照对在 2016年和 2020年均具有成功的 HTTP 状态码。在新增的约 8,500 对配对的网页存档快照中，如图10b所示，约 38% 的快照对在 2016年和 2020年均具有成功的 HTTP 状态码。总计约有 11,000 对 2016/2020 网页存档快照符合索引条件。EDGI 通过使用互联网档案馆的 CDX API，以尽可能减少未成功的 HTTP 状态码的方式收集数据。[17] 对于新增的快照对集合，其他网页档案没有面向公众的 CDX API，而 TimeMap 中也不包含任何状态码信息，因此新增快照对集合中出现未成功 HTTP 状态码的网页更多。

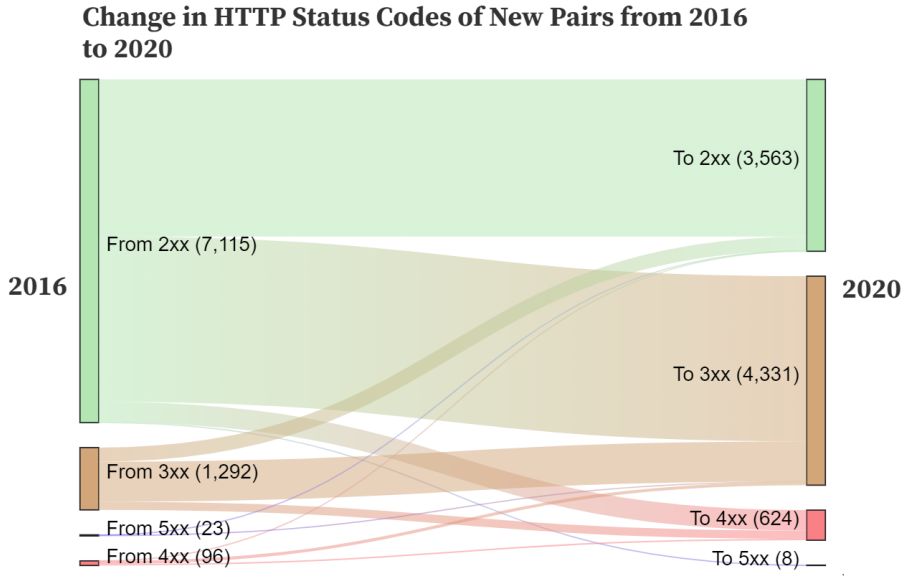
时光回溯机保存了原始 EDGI 数据集中的所有配对网页存档快照。相比之下，新增的配对集合中的网页存档快照分布在多个网页档案中。在这个新集合中，2016年和2020年均具有成功状态码的网页存档快照共有3,563个。表1根据这些快照的时间地图，展示了保存它们的网页档案。部分 URI-R 在所关注的时间范围内（即2016年和2020年的上半年）有多个有效快照。在表中，这些 URI-R 按每个网页档案分别计数。百分比表示至少有一个

[17]https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server

使网页中的变化可被发现



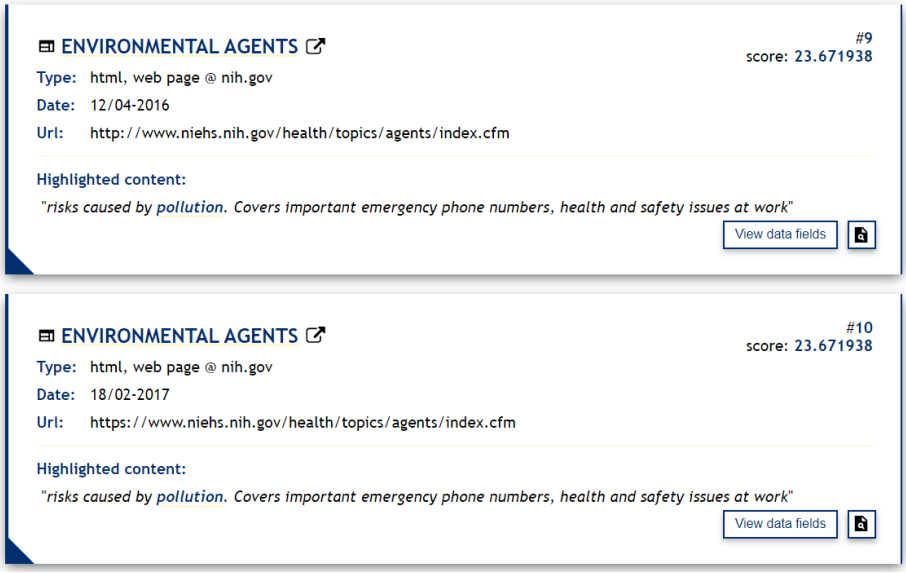
(a) EDGI 识别出的配对网页存档快照中，约 75% 在 2016年和 2020年均具有成功的 (200) HTTP 状态码。



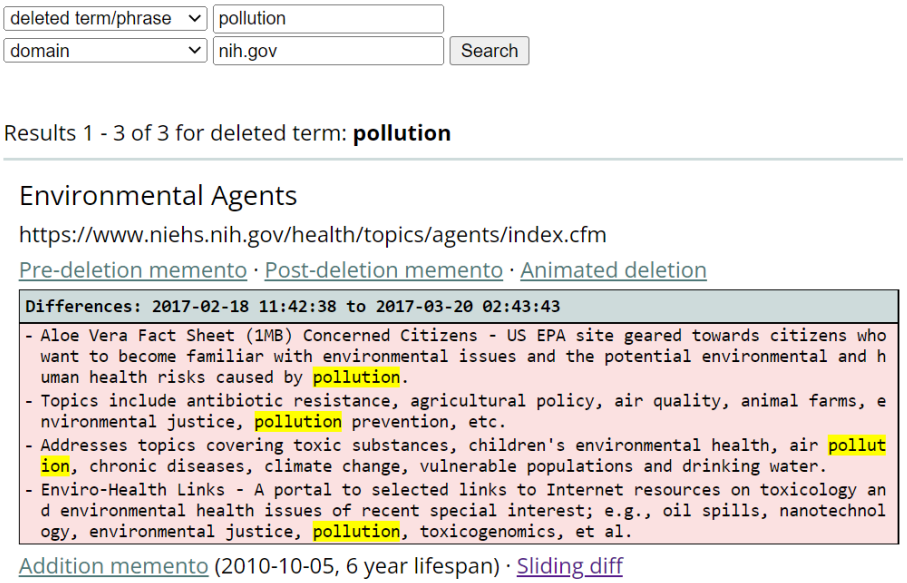
(b) 在　新的配对集中，另有 3,500 组配对网页存档快照在 2016年和 20　20.

图10。在 EDGI 数据集中约40,000个种子 URI-R 中，约有11,000个在2016年和2020年均具有成功（200）HTTP 状态码。我们还从多个网页档案中找到了另外3,500个具有成功状态码的网页存档快照。

使网页中的变化可被发现



(a) SolrWayback 搜索引擎结果页



(b) 变更文本搜索引擎结果页

图 11. 在 SolrWayback 中，同一网页的多个版本会显示在搜索结果中，但不会说明它们之间有何不同。在变更文本搜索引擎的结果页中，清楚标出了网页 https://www.niehs.nih.gov/health/topics/agents/index.cfm 中与查询词“污染”的新增和删除相匹配的版本。

使网页中的变化可被发现

网络档案	%（2016年）	%（2020年）
webarchive.loc.gov	59.81	60.93
web.archive.org	34.41	71.15
wayback.archive-it.org	12.41	10.36
arquivo.pt	5.95	3.59
perma.cc	0.39	0.39
web.archive.org.au	0.11	0.08
swap.stanford.edu	0.11	0.06
wayback.vefsafn.is	0.11	0.20
waext.banq.qc.ca	0.08	0.08
www.webarchive.org.uk	0.03	0.08
archive.md	0.03	0.06

表1. 2016年和2020年上半年包含网页存档快照的网页档案，针对新配对集中状态码为200至200的3,563个网页，如图10b右上方所示。

指定时间范围内在 TimeMap 中找到的 Memento 数量，除以 3,563。百分比便于比较不同存档机构之间以及不同年份的馆藏情况。由于每个 URI-R 在多个存档机构中都可能有 Memento，因此各列的总和没有实际意义。

首先，美国国会图书馆的网页档案保存了许多网页的网页存档快照，这些网页在2016年上半年无法通过时光回溯机访问。未来研究美国联邦政府网站的研究人员应在使用时光回溯机的基础上，同时利用美国国会图书馆的网页档案。其次，Arquivo.pt似乎在2016年5月对美国联邦政府网站进行了抓取，而时光回溯机和美国国会图书馆均没有相近日期时间的网页存档快照。使用MemGator查询汇集所有网页档案中的网页存档快照列表，不仅能增加可供分析的历史证据，还能揭示完整的档案集合及创建这些档案集合背后的可能动机。

对成对的 2016/2020 网页存档快照进行索引，是确定每个网页还需索引哪些其他版本的起点。变更文本计算脚本将确定每个已索引网页的两个版本之间新增和删除的所有词语。随后，可以对该网页的其他版本进行二分搜索，从而提高确定词语变更时间的时间粒度。EDGI 已识别出含有词语或短语删除的网页，是优先进行索引的理想候选。由于 EDGI 只追踪了约 50 个词语和短语，因此可以从变更文本计算结果中识别出其他有意义的词语。

初始索引仅包含一小组数据，共 100 对网页存档快照。EDGI 计算得出，这些快照对中，术语“可持续性”“污染”“人为的”或短语“濒危物种”遭到了全部或部分删除。下一轮索引规模更大，包含 1,000 对，即 EDGI 匹配对总数的 10%。这些快照对中，术语和短语“有毒的”“清洁能源”“气候变化”和“全球变暖”遭到了全部或部分删除。初始小样本中显现的许多趋势在较大样本中依然存在，因此这些趋势很可能贯穿整个数据集。

使用 Hypercane 下载 1,000 对数据（图6中第 1 级的一部分）耗时 41 小时。使用 UKWA WARC索引器对 WARC 文件进行初始 Lucene 索引耗时 2.5 小时，PyWB 索引耗时 40 分钟。变更文本计算脚本生成要发布到 Lucene索引的 JSON 更新数据，耗时不超过 10 秒；发布数据耗时不超过 5 秒。这些步骤对应图6中的第 2 级。

5.2 讨论

与其他面向网页档案的时间搜索引擎（如 SolrWayback）以及 EDGI 使用的网络监测策略相比，变更文本搜索引擎结果页面具备更多功能。变更文本搜索引擎会将同一网页的多个版本分组，以便用户理解它们之间的差异，如图11所示。在 SolrWayback 中，同一网页的多个相同版本会显示在搜索结果页面中，而删除之后紧接着的版本则不会出现在搜索结果中。相比之下，变更文本搜索引擎结果页面只显示查询词发生有意义变化的版本。用户还可以借助差异片段查看上下文中的变更，并通过每条结果中的链接使用滑动差异和动画工具查看更多细节。网络监测策略无法用于查看上下文中的变更。

使用 Lucene 对文本内容进行索引，与 EDGI 使用 Python 完成的删除词计算之间，一个主要差异在于两者采用了完全不同的样板文本移除技术。总体而言，Lucene 的样板文本移除会比 EDGI 的技术从网页中移除更多内容。这意味着，有些词项被 EDGI 识别为已删除，但在 Lucene 索引中却并非如此。由于样板文本移除的缘故，Lucene 对其他一些网页的索引效果较差，这是删除词计算中的另一项差异。EDGI 识别为已删除的词项中，有些位于各种导航页面部分，这些部分在样板文本移除过程中并未被移除，因此是否应将这些词项称为删除内容，取决于个体研究者对样板文本移除的偏好。此外，使用 Lucene 对所有内容进行索引，可以获取所有删除词项；而 EDGI 的网页监测技术只能追踪预定义词。无需预先定义即可查找删除词项的能力十分强大，对某些研究者而言，这或许足以弥补索引效果不佳的问题。

虽然最初的 EDGI 研究仅追踪了 56 个具有环境动机的词项和短语，但如今还能发现其他常见的删除词项，如表2所示。最初的词项列表包括“安全”“透明度”“监管”和“工作”等词项，但“公众”“获取”“行动”“发展”和“科学”也是含义相近且同样常被删除的词项。许多排名靠前的删除词项是停用词或时间词。最初 EDGI 词项中的两个词——“change”和“state”——是停用词，但在联邦环境数据集中具有特定含义。[18]

按删除次数从多到少排列，新发现的词项包括：国家的、支持、公众、计划、资源、过程、数据、包括、美国、发展、联合的、学习，

部门 ，行动 ，获取 ，work，影响 ，工具 ，
领域、搜索、实验室、技术、努力、包括、自然、科学、规划、处理和开放。

术语类型	示例	数量
停用词	about, more, which, state	49
时间类	2015，2，一个，年	13
EDGI	气候，清洁，影响，水	11
新发现	公众，发展，获取，科学	29

表2。对 1,000 个配对备忘录索引样本中排名前 100 的删除词项进行分类。

[18]https://countwordsfree.com/stopwords

通过检查新删除词项的搜索结果，还可以进一步发现删除的短语。例如，“公众”是一个新的删除词项，与原始词项“透明度”类似。人工检查“公众”的搜索结果后发现，该词经常出现在被删除短语的语境中。“公众意见”是删除词项“公众”的常见语境，这种用法与原始词项“透明度”类似。另一个更常见的语境是短语“公共卫生”，它实际上与原始词项“安全”的相似度高于“透明度”。

这两种用于详细检查变化的工具——滑动差异和动画——各自在不同场景中具有优势。滑动差异工具可以显示同一网页的两个以上版本，而动画一次只能显示两个版本。滑动差异还会持续显示变化，而动画视图中的已删除文本则会自然消失。滑动差异工具的另一项优势是加载速度极快，因为它只比较内容。动画则必须通过重放加载两个网页存档快照、计算 HTML 差异，然后加载并显示动画，因此耗时要长得多。滑动差异能很好地显示文本块随时间发生的变化，但差异视图会隐藏未发生变化的内容。如果用户想在上下文中查看整个网页，动画会显示完整页面，包括未发生变化的文本和图像。与滑动差异工具相比，动画也能更突出地呈现差异。当网页上的变化过多时，这两种工具都会受到影响，因为高亮内容过多，而且动画运行过慢。

EDGI 联邦政府网页数据集对于评估这一变更文本搜索引擎既有价值，也很合适。政府创建的文件不享有隐私权；事实上，这些文件应当被存档，并让公众能够检索到。然而，研究人员在为网络档案集合建立索引时必须格外谨慎。异质的网络档案集合可能包含应受到更严格隐私保护的材料，例如未成年人建立的网站。Lin 等人 [22] 推测，人们不知道自己的网站被存档了，因为他们从未明确同意；而其他人则依赖缺乏网页存档搜索功能来保护隐私。Mackinnon [23] 在一项用户研究中证实了这两种假设。Lin 等人和 Mackinnon 都认为，匿名化可作为一种工具，使研究人员能够研究整体网络档案集合，同时不损害内容创作者的隐私。

6 结论与未来工作

现有的网页档案时间搜索引擎不允许用户查询随时间发生的变更。本文介绍了一种变更文本搜索引擎，用户可以借此查找并查看网页中的变更。搜索引擎结果页将同一网页的多个版本归为一组，同时不会隐藏这些版本之间的变更。事实上，版本间的变更是能够进行分组的核心原因，因为这种分组体现了某个词项在网页上的存续期。删除动画能够在上下文中展示变更，而滑动差异查看器则可以快速检查网页多个版本之间的差异。倒排索引包含语料库中最常被删除词项的宝贵信息。

未来关于架构的重要工作之一，是实现索引流程自动化。Kiesel 等人 [20] 已为 Lucene 和 PyWB 自动化了个人网络存档的索引。要使该系统正常运行，还需要将变更文本计算脚本自动化。Jackson 等人[16] 认为，实现 Lucene/PyWB 双重索引自动化，是网页存档社区全文搜索架构的一项重大进展。搜索引擎代码的另一项改进，是增强对新增短语的检测能力：即使两个词项已分别出现在网页上，也能识别由它们组成的新增短语；此外，还应识别部分新增术语。

为网页索引更多较早的版本，将有助于深入了解词项何时被新增，例如在之前的总统任期内。较早的版本还将帮助发现以往任期内删除的其他词项和短语。从计算视角查看索引，有助于计算出新的已删除词项。后续工作还将包括计算新的已删除短语。未来工作的另一方面，是对网页的变更程度进行分类。这项工作将有助于区分不同类型的变更，例如为配合组织的新目标而重写整个网页，或删除部分内容区块，从而使公众无法获取重要信息。

通过考察具体的用户任务，可以证明对变更文本搜索引擎以及特定搜索引擎结果页功能的需求是合理的。Jayanetti 等人 [18] 对网页档案中的用户进行了分类，包括查看同一网页多个版本的用户。未来关于用户界面的工作将包括分析这些用户的任务，以更好地确定这些用户的需求。未来的工作还将包括对搜索界面前端（包括动画）进行用户研究，以评估记者等用户能否更成功地完成其信息检索任务。

参考文献

[1] Samantha Abrams、Alexis Antracoli、Rachel Appel、Celia Caust-Ellenbogen、Sarah Denison、Sumitra Duncan 和 Stefanie Ramsay。2019。为打造更易用的网页档案播下种子：一项关于 Archive-It 可用性的研究。《美国档案工作者》82，2（2019），440–469。[2] Eytan Adar、Mira Dontcheva、James Fogarty 和 Daniel S. Weld。2008。Zoetrope：与易逝网络交互。载于第*21*届 *ACM* 用户界面软件与技术研讨会论文集（美国加利福尼亚州蒙特雷）（*UIST '08*）。计算机协会，纽约州纽约市，美国，239–248。https://doi.org/10.1145/1449715.1449756[3] Eytan Adar、Jaime Teevan、Susan T Dumais 和 Jonathan L Elsas。2009。网络改变了一切：理解网页内容的动态变化。载于第二届 *ACM* 网页搜索与数据挖掘国际会议论文集。282–291。[4] Sawood Alam 和 Michael L. Nelson。2016年。MemGator——便携式并发 Memento 聚合器：Go 语言跨平台 CLI 和服务器二进制文件。载于第*16*届 *ACM/IEEE-CS* 数字图书馆联合会议论文集（美国新泽西州纽瓦克）（*JCDL '16*）。计算机协会，纽约州纽约市，美国，243–244。https://doi.org/10.1145/2910896.2925452[5] Klaus Berberich、Srikanta Bedathur、Thomas Neumann 和 Gerhard Weikum。2007。一台文本搜索时间机器。载于第*30*届 *ACM SIGIR* 信息检索研究与发展国际会议论文集（荷兰阿姆斯特丹）（*SIGIR '07*）。计算机协会，纽约州纽约市，美国，519–526。https://doi.org/10.1145/1277741.1277831[6] Jack Cherng 和 Chris Boulton。2023。*jfcherng/php-diff*。https://github.com/jfcherng/php-diff[7] Fanny Chevalier、Pierre Dragicevic、Anastasia Bezerianos 和 Jean-Daniel Fekete。2010。运用文本动画过渡支持文档历史记录导航。载于*SIGCHI* 人机交互系统中的人为因素大会论文集（美国佐治亚州亚特兰大）（*CHI '10*）。计算机协会，纽约州纽约市，美国，683–692。https://doi.org/10.1145/1753326.1753427[8] Laurent Denoue、Scott Carter 和 Matthew Cooper。2018。SlideDiff：为幻灯片中的文本和媒体变化制作动画。载于*2018*年 *ACM* 文档工程研讨会论文集（加拿大新斯科舍省哈利法克斯）（*DocEng '18*）。计算机协会，纽约州纽约市，美国，文章37，共4页。https://doi.org/10.1145/3209280.3229107[9] Thomas Egense、Toke Eskildsen 和 Anders Klindt Myrvoll。2022。*SolrWayback*。https://github.com/netarchivesuite/solrwayback[10] Lesley Frew。2022。大众媒体中的网页存档 II：记者的用户任务。https://ws-dl.blogspot.com/2022/08/2022-08-04-web-archiving-in-popular.html[11] 国际互联网保存联盟访问工作组。2006。互联网档案访问用例。技术报告。https://digital.library.unt.edu/ark:/67531/metadc1457746/[12] Christopher G. Healey、Kellogg S. Booth 和 James T. Enns。1996。运用前注意加工进行高速视觉估算。*ACM* 人机交互汇刊3，2（1996年6月），107–135。https://doi.org/10.1145/230562.230563[13] Austin Z. Henley 和 Scott D. Fleming。2016年。Yestercode：改进可视化数据流编程环境中的代码变更支持。载于*2016*年 *IEEE* 可视化语言与以人为中心的计算研讨会（*VL/HCC*）。106–114。https://doi.org/10.1109/VLHCC.2016.7739672[14] Andrew Jackson。2022。网页档案发现 - *WARC* 索引器。https://github.com/ukwa/webarchive-discovery/tree/master/warc-indexer[15] Andrew Jackson、Jimmy Lin、IanMilligan 和 Nick Ruest。2016年。支持学术活动的网页档案探索式搜索界面设计愿景。载于第*16*届 *ACM/IEEE-CS* 数字图书馆联合会议论文集（美国新泽西州纽瓦克）（*JCDL '16*）。计算机协会，纽约州纽约市，美国，103–106。https://doi.org/10.1145/2910896.2910912[16] Andrew Jackson、Anders Klindt Myrvoll、Toke Eskildsen、Thomas Egense 和 Ben O'Brien。2022。网页档案全文搜索。载于*IIPC* 网页存档会议。https://netpreserve. org/ga2022/wac/abstracts/。[17] Adam Jatowt、Yukiko Kawai、Satoshi Nakamura、Yutaka Kidawara 和 Katsumi Tanaka。2006。Journey to the Past：历史网页浏览器框架提案。载于第十七届超文本与超媒体会议论文集（丹麦欧登塞）（*HYPERTEXT '06*）。计算机协会，

使网页中的变化可被发现

纽约州纽约市，美国，135–144。https://doi.org/10.1145/1149941.1149969[18] Himarsha Jayanetti、Kritika Garg、Sawood Alam、Michael L. Nelson 和 Michele C. Weigle。2022。网页档案中的机器人数量仍超过人类，但差距有所缩小。载于数字图书馆理论与实践会议论文集（TPDL）。[19] Shawn M. Jones、Michele C. Weigle、Martin Klein 和 Michael L. Nelson。2021年。Hypercane：对网络档案馆藏进行智能抽样。载于2021年 ACM/IEEE数字图书馆联合会议（JCDL）。316–317。https://doi.org/10.1109/JCDL52503.2021.00049[20] Johannes Kiesel、Arjen P. de Vries、Matthias Hagen、Benno Stein 和 Martin Potthast。2018年。WASP：个性化网页归档与搜索。载于第一届实验搜索与信息检索系统设计双年会议论文集，意大利贝尔蒂诺罗，2018年8月28日至31日（CEUR研讨会论文集，第2167卷），Omar Alonso 和 Gianmaria Silvello（编）。CEUR-WS.org，16–21。http://ceur-ws.org/Vol-2167/paper6.pdf[21] Ilya Kreymer。2022。PyWB——人人可用的网页归档工具。https://github.com/ikreymer/pywb[22] Jimmy Lin、Ian Milligan、Douglas W. Oard、Nick Ruest 和 Katie Shilton。2020年。我们能做到，但应该做吗？提供访问 GeoCities 及其他历史数字馆藏时的伦理考量。载于2020年人类信息交互与检索会议论文集（加拿大不列颠哥伦比亚省温哥华）（CHIIR '20）。计算机协会，纽约州纽约市，美国，135–144。https://doi.org/10.1145/3343413.3377980[23] Katherine Mackinnon。2022。Databound：在万维网上成长的历史。博士学位论文。多伦多大学（加拿大）。[24] Daniela Major。2023。众人拾柴，方能建成档案。https://webarchivingrt.wordpress.com/2023/01/18/it-takes-a-village-to-raise-an-archive-how-the-use-of-web-sources-fosters-collaboration/[25] Fernando Melo、Hugo Viana、Daniel Gomes 和 Miguel Costa。2016年。葡萄牙网络档案搜索系统第二版的架构。技术报告。葡萄牙网络档案 Arquivo.pt。https://sobre.arquivo.pt/wp-content/uploads/architecture-of-the-portuguese-web-archive-search-1.pdf[26] Eric Nost、Gretchen Gehrke、Grace Poudrier、Aaron Lemelin、Marcy Beck、Sara Wylie 代表环境数据与治理倡议。2021年。可视化美国联邦环境机构网站在2016年至2020年的变化。PLOS ONE 16，2（2021年2月），1–27。https://doi.org/10. 1371/journal.pone.0246450[27] Jessica Ogden、Ed Summers 和 Shawn Walker。2021年。使用模式：对网页档案在在线话语中作用的概念化。https: //www.resaw2021.net/ 第四届 RESAW（档案网页材料研究基础设施）会议：网络史与网页档案中的主流与边缘内容，RESAW21；会议日期：2021年6月17日至2021年6月18日。[28] Arnold Overwijk、Chenyan Xiong 和 Jamie Callan。2022。ClueWeb22：包含丰富信息的100亿份网页文档。载于第45届 ACM SIGIR 信息检索研究与发展国际会议论文集。3360–3362。[29] Ignacio Perez-Messina、Claudio Gutierrez 和 Eduardo Graells-Garrido。2018年。文档演变的有机可视化。载于第23届智能用户界面国际会议（日本东京）（IUI '18）。计算机协会，纽约州纽约市，美国，497–501。https: //doi.org/10.1145/3172944.3173004[30] Mark Phillips、Dan Chudnov 和 James Jacobs。2016年。对任期结束网页档案的探索性分析：两个档案集合的比较。《网页归档与数字图书馆》，JCDL 2016（2016）。[31] Mark E Phillips 和 Kristy K Phillips。2017。2016年任期结束总统网页档案。《Against the Grain》29，6（2017），10。[32] Marcel Ras 和 Sara van Bussel。2007。网页归档用户调查。技术报告。荷兰国家图书馆（荷兰皇家图书馆）。[33] Tracy Seneca、Abbie Grotke、Cathy Nelson Hartman 和 Kris Carpenter。2012。众人拾柴，方能保存网络：任期结束网页档案。《Documents to the People》（DttP）40（2012），16。[34] Tim Sherratt 和 Andrew Jackson。2022。GLAM-Workbench/web-archives。https://doi.org/10.5281/zenodo.6450762[35] David Smiley、Eric Pugh、Kranti Parisa 和 Matt Mitchell。2015。Apache Solr 企业搜索服务器。Packt Publishing Ltd.[36] Jaime Teevan。2007。“它去哪儿了？”：人们如何询问丢失的网络信息。《美国信息科学与技术学会会刊》44，1（2007），1–19。[37] Herbert Van de Sompel、Michael Nelson 和 Robert Sanderson。2013。RFC 7089——基于时间获取资源状态的 HTTP 框架——Memento。技术报告。https://tools.ietf.org/html/rfc7089[38] Dakuo Wang、Judith S. Olson、Jingwen Zhang、Trung Nguyen 和 Gary M. Olson。2015。DocuViz：协作写作的可视化。载于第33届 ACM人机交互系统年度会议论文集（韩国首尔）（CHI '15）。计算机协会，纽约州纽约市，美国，1865–1874。https://doi.org/10.1145/2702123.2702517