基于数据挖掘的Web Archive资源应用分析*

吴振新[1] 张智雄[1] 孙志茹[1,2]

[1]（中国科学院国家科学图书馆，北京 100190）[2]（中国科学院研究生院，北京 100049）

[摘要] Web archive 中蕴含着丰富的信息资源，是一个有待开发的巨大信息宝库。目前人们对 Web archive 资源的应用已进行初步尝试和努力，并取得一定进展。本文概括地介绍了 Web archive 资源应用的基本情况，而后从数据挖掘的角度，对 Web archive 资源的深层次应用进行总结和分析。

[关键词] Web archive，应用分析，数据挖掘

An Analysis of the Application of Web Archive Resources Based on Data Mining

Wu Zhenxin

(The Library of Chinese Academy of Science，Beijing 100190,China)

Sun Zhiru

(The Library of Chinese Academy of Science，beijing 100190,China)

(Graduate School of the Chinese Academy of Sciences,beijing:100049,China)

[Abstract] Web archive , which contains a wealth of information resources to be developed ,is a great treasure-house of information. Researchers have made efforts to try to use archived resources and achieve some progress. This article introduced current applications of web archive resources, and then from the perspective of data mining, analyzes and sums up the in-depth applications of web archive resources.

[Keywords] Web Archive　Application Analysis　Data Mining

1 引言

网络信息资源保存（Web archive）在不同的时间结点对网页进行持续地保存，这些网页不只是被收集存储起来，而且保持了搜集时间和原有的链接关系，它按照时间轴的发展，真实地记载了网络的变化，保存了社会的变迁和文明（文化）的发展，Web archive资源已经成为一笔不可或缺的巨大社会财富。

经过多年的发展积累，目前Web archive的存储量已经非常可观。从 1996 年开始进行全球网页保存的Internet Archive[1]（简称IA)至今已保存了 850 亿个页面。“中国Web信息博物馆”[2](简称WebInfomall)目前已经维护有自 2001 年以来采集和保存的 30 亿以中文为主的网页，并以平均每月四千五百万网页的速度扩大规模。如何利用和挖掘这个信息宝藏，成为Web archive研究当中非常重要的课题。

2 Web Archive 资源应用研究现状分析

目前全球近百个项目进行了 Web Archive 的研究和实践，Web Archive 资源

* 本文受国家社会科学基金项目“网络信息资源保存的理论与方法研究”课题的资助，课题编号为 06BTQ025。

在不断地积累和膨胀，但绝大多数项目仅对采集来的资源提供基本的浏览和检索，只有少数项目利用数学、统计、随机过程分析等手段，结合信息分类、数据挖掘、计算语言学的有关技术，对上千万量级的信息从不同层次进行开采和提炼，对 Web Archive 资源的利用进行了更深层次的研究。

目前主要包括这样几种应用：

(1) 网站重现

网站重现是将Web archive存储的网站内容以其原有的样貌展现给用户的过程。网站重现最简单、最普遍的应用是为用户提供存档内容的访问和浏览。基本原理是利用重现工具从Web archive的仓储库（repository）中获取页面并尽可能忠实地显示出来，让用户感觉就像是在访问原始网站一样。为此,各项目开发了不同的重现工具，比较典型的有IA的开源“网站时光倒流机”Wayback Machine[3]；日本京都大学提出了一套对现有的和历史网页进行组合浏览的工具[4]；IIPC提供的开源软件WEAR[5]；加利福尼亚数字图书馆的开源索引与查询工具XTF[6]；澳大利亚国家图书馆开发的一款XML Archive查询工具Xinq[7]等。另外,澳大利亚国家图书馆、Web infomall等所使用的Web Archive系统都实现了网页的重现。

除此之外，网站重现还有以下两个方面的具体应用：

（某个时间结点的）网站恢复。IA就以其采集并存储的网页信息帮助很多网站做站点恢复。美国Old Dominion 大学采用Warrick[8]通过lazy preservation[9]的方法，递归地在 4 个web仓储库中爬行来帮助重建和恢复网站，这 4 个web仓储库包括IA的历史存档库和Google、Live Search和Yahoo这几个搜索引擎的缓存库。

英国国家档案馆提出了一个“网络连续性”（Web Continuity[10]）的创新项目，该项目对所有重要的政府网站提供归档和重定向服务。这些政府网站的用户在遇到浏览器的 404 （找不到档案）错误信息时，将被自动重定向到对应的存档网页。该项服务通过提供无缝的导航和帮助网站进行自动归档功能的整合，极大地提高了用户的体验。

(2) 对于 Web 自身的研究

Web archive不仅拥有网页内容本身，还拥有很多与Web相关的技术信息，如Web结构、数量、链接等等。目前的研究基本从三个不同的层次展开，即信息资源量及其分布；海量信息之间的关联结构；信息内容。

Web informall基于2001年6月以来所做的几次大规模搜集、组织与整理的中国Web数据，对中国网页数量、平均大小、域名内的网页数量、平均大小、域名分布等进行了初步统计。同时通过系统搜集过程中记录的Web链接结构信息分析了中国Web的大小、形状和结构，包括Web直径、幂率、社区。此外还发现了网络信息量的成长规律即增长指数，构建了计算网页生命周期的模型，同时对从海量网页信息中识别网络社区进行了探索。这些研究进一步揭示了中国Web的特性。

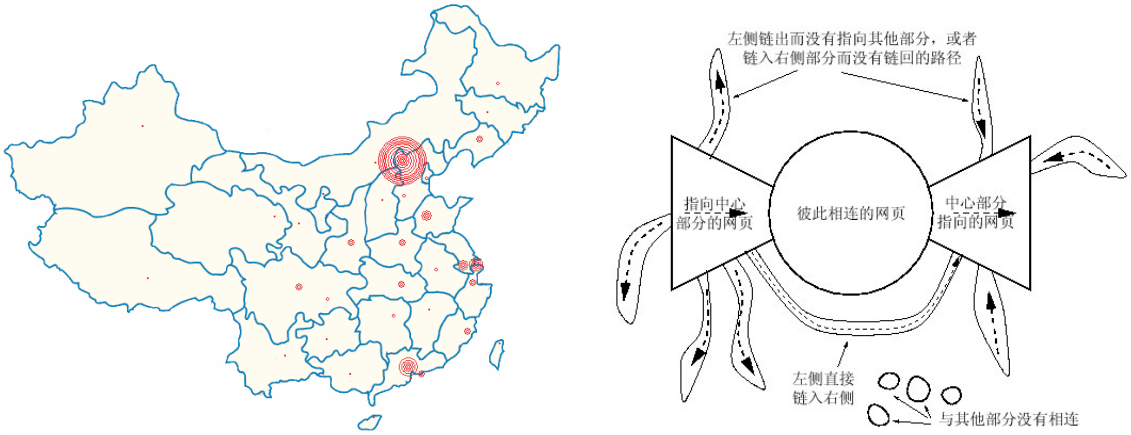

图1 中国各省市网站分布（左）以及从网页结构推断中国Web的形状（右）[11]

澳大利亚在线存档处理项目AOLAP采用数据仓库技术处理所获取的大量数据，并利用所获取的数据来研究Web，分析的内容包括：Web服务器的分布、域名分布、文件类型在不同Web服务器的分布情况等内容。利用这些数据还可以进一步分析澳大利亚网络空间的现状、应用情况、Web保存面临的问题等[12]。

另外通过对Web的分析还可以为Web archive决策提供支持作用。例如，分析不同类型、不同领域、不同专题网站的复杂程度、更新频率、域（.org, .com.）、文件类型比例（.pdf, .html）等可以帮助Web archive设定爬行策略；分析Web技术的演化可以作为Web archive项目中技术选择的基础[13]。

(3) 基于数据挖掘的 Web Archive 资源应用

这种应用已经从简单的数据统计分析过渡到信息分析、从数据的处理过渡到知识发现。主要是利用数据挖掘、自然语言处理等知识技术对 Web Archive 所保存的海量信息进行深层次分析和研究，从中获取隐含在其中的有用知识。

3 基于数据挖掘的 Web Archive 资源应用分析

基于数据挖掘的 Web archive 资源应用是一个非常具有挑战性的研究，虽然还是刚刚起步，但已经展现出广阔的前景和巨大的潜力。利用这些信息可以研究 Web 本身的历史演化、网络社区形成（网络社区可以定义为由于共同利益或兴趣而自发联系起来的网页）、专题网络爬行、思想和实践（行为）的传播、专题事件历史追踪、社会关系分析等等。目前对于 Web Archive 资源挖掘主要应用于社会科学领域。

然而使用网络数据进行社会科学研究也面临相当大的挑战：

(1) 内容和结构是隐含的，而不像传统的表格数据库有具体字段和关系。需要新的工具把数据解析转化为有意义的成分和结构。

(2) 来自许多代理人（包括个人和组织）的独立运行的数据结果，要远胜于出于特定目的集中收集数据。因此，研究范式从数据收集和分析转变为数据挖掘和信息发现。

(3) 规模庞大、数十亿的数据数量带来存储和计算上的挑战。伴随着 PB 级存储以及数据密集型计算能力的不断发展，未来两年在技术上是可行的。但人工无法处理如此大量的数据，需要新的工具辅助，如半监督的机器学习。

(4) 即使是采集网络上公开提供的数据，隐私依然是一个重大问题。通过分析和挖掘相结合，可以很容易地揭示一些原始来源中并不明显的信息。因此，必须采用基于隐私保护的数据挖掘和知识发现技术发现。

3.1 基于Internet Archive的数据挖掘——Web Library[14]

2005 年美国 Cornell 大学的社会科学学院在国家科学基金的资助下开展了“面向社会科学研究的超大规模半结构化数据集合”的研究，以 IA 的网页仓储为基础构建了面向社会科学的 Web Library，开展了一系列基于计算的社会学研究，其主要目的是支持社会学和历史学研究人员利用 Web 数据进行研究工作。

Web Library 的数据来源于 IA，这些数据都是采集后未被加工的原始数据。系统将这些数据通过 Internet2 传输到 Library 并存储在 tape archive 中。然后预载子系统（Preload System）将输入的 ARC 文件和 DAT 文件解压缩，解析这些文件来抽取元数据，并生成两种类型的输出文件：用于载入数据库的元数据和将要被存入页面存储（Page Store）的 Web 页面内容。

数据库（Database）用来存储每个 Web 页面的元数据，页面存储（Page Store）提供了每个唯一 Web 页的单独副本。Tape Archive 中装有所有从 IA 获得的数据。Tape Archive 与 Database 一起，最终提供相当于 IA 的部分 Web 内容的索引镜像。

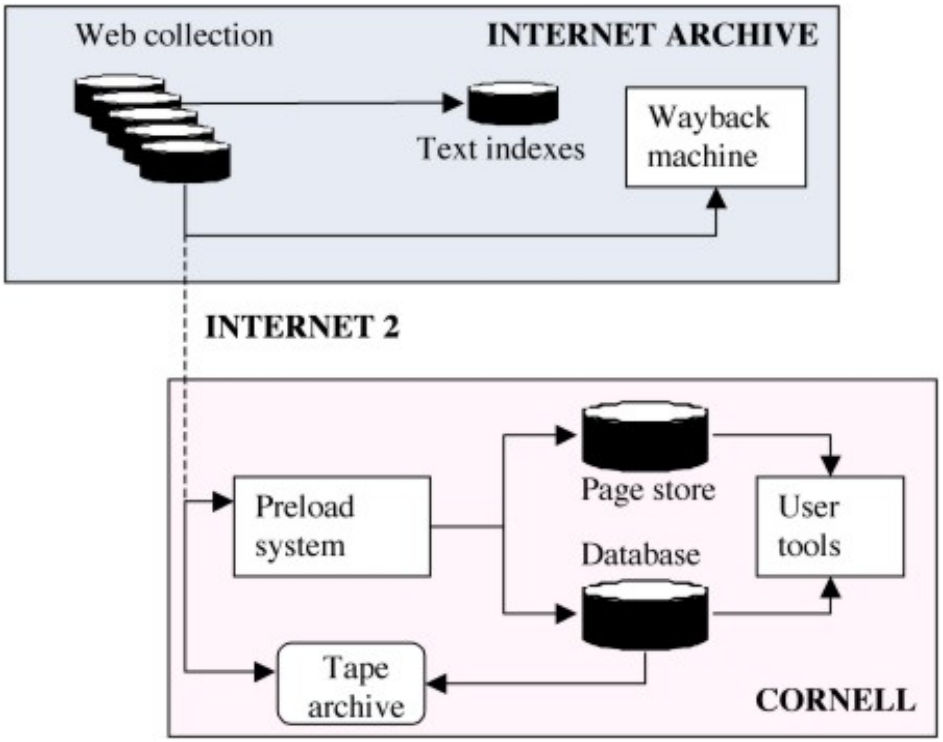

图 1 Web Library的数据流和数据存储[15]

该项目对所在学院的研究人员（社会学、通信学、信息科学）进行调查发现：

几乎所有研究人员都趋于使用较小规模、更易于管理的 Web 数据集而不是完整的数据集。因此项目为研究人员开发了两种类型的工具：一种用于从完整集合中抽取子集并建立索引，另一种用于较小子集的分析。

目前 Web Library 支持的用户服务可分为三类：1）基本访问服务，系统提供 Web Service API 让客户通过页面来访问集合中的任何元数据；2）子集抽取服务，支持下载部分数据到用户自己的计算机上做进一步分析。3）允许技术水平高的用户在中央计算机上运行自己的程序。

Web Library 有计划地开发了一系列处理 Web archive 子集的用户工具，并准备利用自然语言处理和机器学习等技术开发更高级的分析工具。目前有 1)“Retro Brower”，允许用户浏览某一特定日期的 Web 网页。2）“Subset Extraction and Manipulation”是一个子集抽取服务工具。3）“Web Graph”是从某次爬行的子集中抽取 Web graph 的工具，用于图形计量的计算，如 PageRank ，或 hub 和认证。4）全文索引是 Web library 面临的最大技术挑战，目前采用 NutchWAX 生成数据集全文索引。

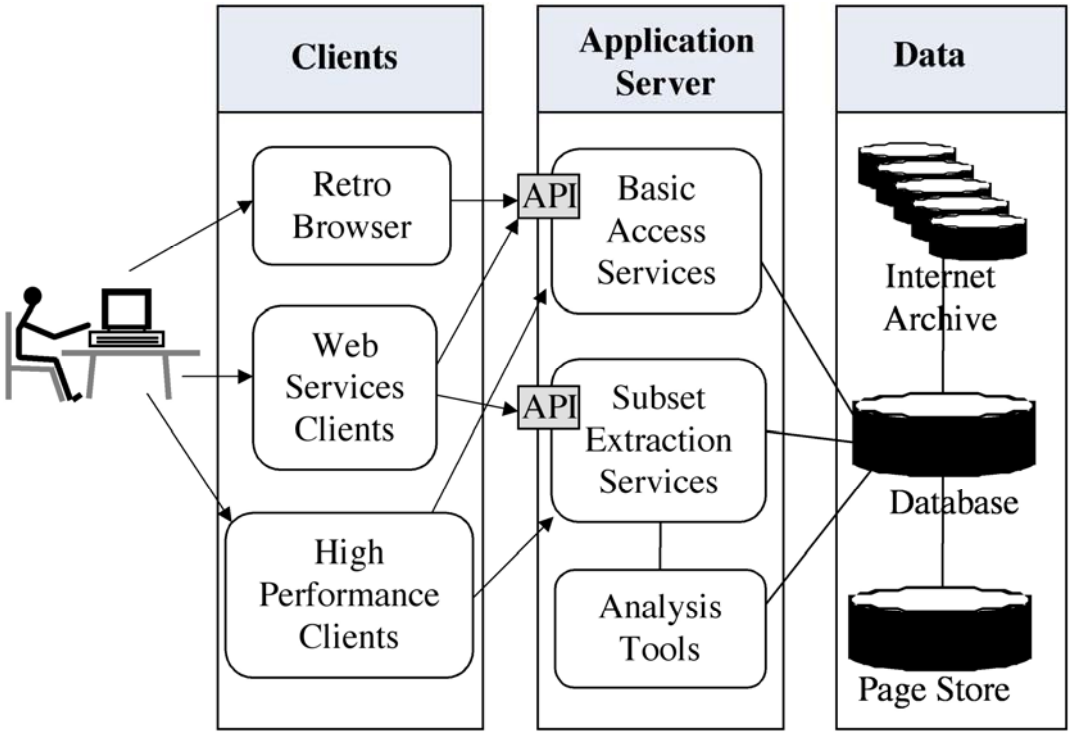

图 2 Web Library用户使用示意图[16]

目前 web library 已经在以下研究中为社会科学研究人员提供了支持：

(1) 专题网络爬行

通过抽取数据子集的工具，用户可以进行专题网络爬行。它采用了一组用于寻找预定主题的相关网页的技术，包括采用了信息检索中的统计方法进行专题爬行；采用了基于监督的机器学习方法来进行资料选择和目录编辑。Web Library 为专题网络爬行提供了一个巨大稳定的测试集，把实验脱离了变幻莫测、无法测

量的实际网络，使专题爬行的实验方法、技术选择和结果可以不断地得以修正。

(2) Web 的结构和演变

计算机专家通过分析网页间的链接图来研究 Web 的结构和演变。Web Library 中非常清晰地存储了数据库中每个页面的全部链接，它可以直接地指明一个子集，如一个 web 快照的全部页面，并且建立显示这些页面间所有链接的邻居矩阵。另外可以通过短语的出现频次来识别新出现的主题。这种方法可以突出在某个时间点网络上正在迅速改变的内容，提供了一种手段来分析新兴媒体的内容，如博客。

(3) 思想和实践（行为）的传播

社会学家特别感兴趣的是研究思想和行为在社会领域中的传播，这对于如社会学、人类学、地理学、经济学、组织行为学、人口、生态和通信等许多学科都是非常重要的。这些研究通常是通过一组特征来识别一个概念，如教育领域中网页上的短语"affirmative action"，该项目通过主题爬行来获得一组具备这些特征的网页，然后通过自动化的工具对其进行分析。

(4) 网络社会变化

与传播研究相似：研究人员通过提取一组网页和使用自动化的工具来分析它们。例如，研究各机构之间的高层管理人员的流动通常会选择 500 家美国最大企业网页作为研究集合，然后利用机器学习方法从新闻稿、财务报告以及类似的文件中抽取管理人员的名字，并跟踪随着时间的变化而发生人员的变化和流动。

3.2 基于Web Archive的社会感知系统研究[17]

网络中的信息来自公司、政府、团体和个人，在现实世界中的各种事件通常都被迅速反映在网络上。认识网络空间的结构并追踪它的变化将使我们能够预先把握先兆，并深刻了解真实社会现象的背景。这种洞察力，不可能通过现有的提供平面（横向）事实的搜索引擎实现。

社会感知系统是一个基于 Web archive 网络信息进行社会行为分析的系统，由日本东京大学工业科学和计算机技术研发中心合作开发。它利用持续了 9 年的以日本域为中心的 Web archive 资源，其基于时间的 URL 索引既能提供对于 URL 的历史（纵向）追踪，同时也提供任意时间点的全部 URL 的（横向）集合。

该系统把 web archive 与分析簇（Cluster）紧密关联，通过并行扫描机制，从 Web Archive 中抽取内容，在负载均衡的基础上，分派给分析簇的其中一个结点进行特定内容处理，并通过一个 5k x 3k 像素的显示墙来更好地进行复杂结果的可视化显示。

(1) 网络结构分析（Web Structural Analysis）

主题相关的网页往往链接着大量的相关超链接，并且在网络拓扑图中位于邻近的位置。利用这一特性，研究人员通过抽取密集子图获得相关网页，并把这些

网页作为一个网络社区。从同一类行业的公司网页到拥有相同爱好的个人网页，在各个领域中都能够形成相应的网络社区。

各社区在网络图中稀疏的部分相互关联着，形成了一个以社区作为结点、其各社区之间关联程度作为边界的关系图，这种图被称为社区图表（community chart），通常被作为网络空间地图。

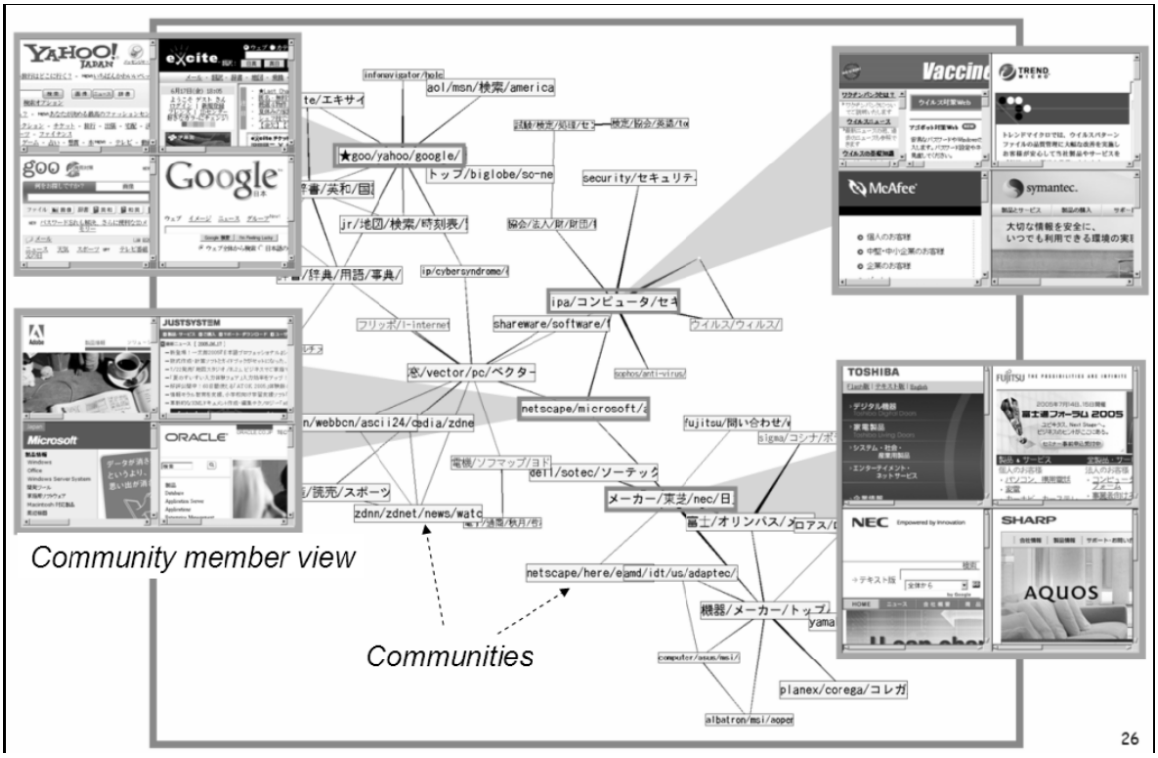

图 3 与“电脑”相关的社区图表子图[18]

图 3 是一个“电脑”的社区图表子图，每个矩形代表了一个社区，相关的社区互相联系着，成员网页也显示出 4 个社区：计算机硬件供应商社区、软件供应商社区、安全信息/供应商社区和门户/搜索引擎社区（由右底到左上）。可以看出这是行业间的关系在网络上的揭示。

另一个非常重要的挑战是分析网页本身文字，如声誉抽取。网站文字内容中包含着源于消费者关于产品和公司的声誉和评价，这对市场营销是非常有益的。然而，提取声誉需要（大量的）详尽列出包含着情绪的词汇（单词和词组），手工构建一个这样的词汇集合是非常昂贵甚至是无法完成的。为此，研究人员采用语言模式和统计方法从 Web archive 中自动建立词汇集合，并开发了声誉提取工具，每个声誉查询从 Web archive 抽取结果并显示给用户，除了原始的文本和正/负声誉，还提取了相关的主题内容。

(2) Web 时序分析（Web Temporal Analysis ）

通过抽取不同时期存在的社区图表，可以追踪相同社区的演化过程。不同时期社区之间的链接可用成员 URL 集合作为识别的标志。随着时间的推移，一些网址加入和离开一个社区，有时也会出现社区的分裂和合并。如果某个时期缺乏

一个社区的对应社区，则意味着一个新社区的出现或一个已有社区的消失。

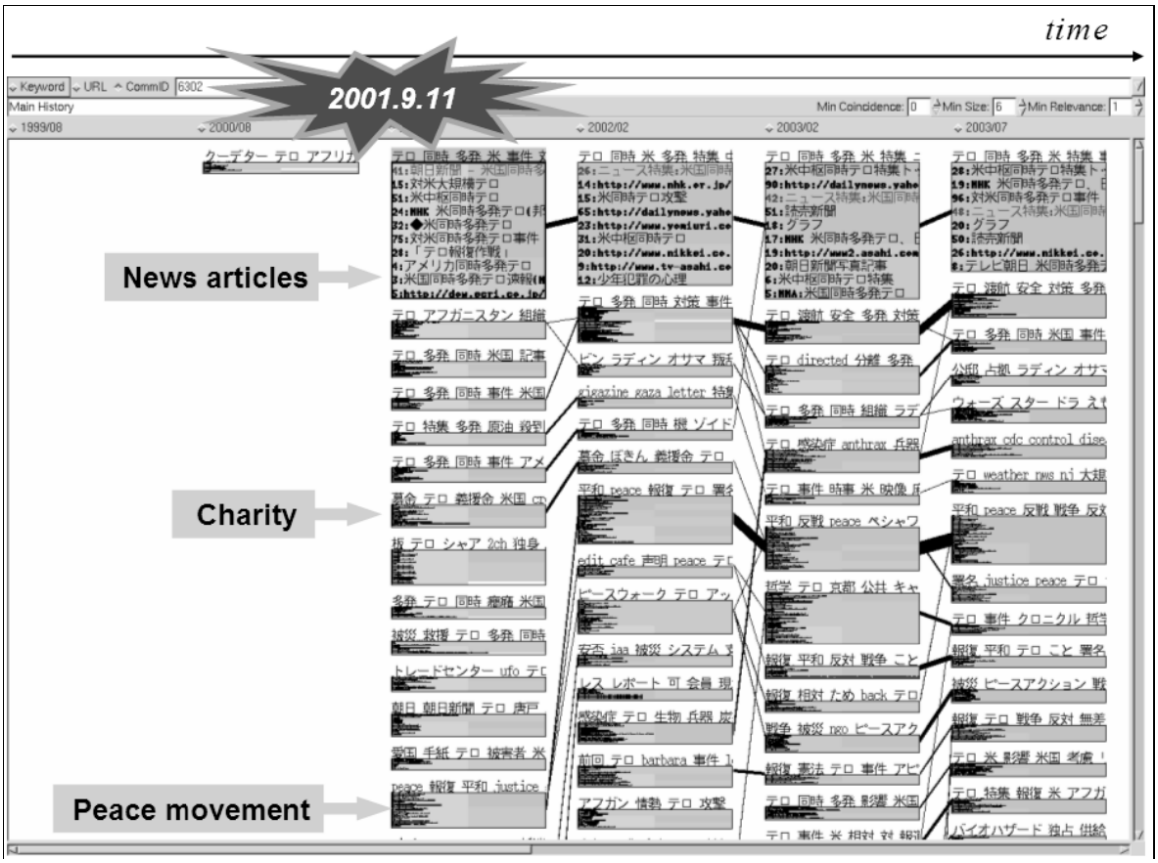

图4. 与恐怖活动有关的社区演变[19]

图 4 中显示了一个社区的可视化演变过程。每栏对应于不同时间，每个矩形代表一个社区，其成员 URL 显示在其内部。相邻时段栏中各社区间的链接被描述而不是链接到每一个时间段。这个例子揭示了在 9.11 之后，与恐怖活动有关的社区突然出现。

该方法可用于调查新信息、主题转换和社会学现象的出现，如用来进行结构图本身演变的可视化，可以观察到社区在萌芽阶段和发展阶段结构图的不同特点。另外在网络上每天都有很多词汇出生和死亡，可采用支持向量机（ SVM ）从网站上提取新的动词、形容词，用于统计和分析新词的变化。

⑶ **消费者行为分析（Consumer Behavior Analysis ）**

博客（Blog）的流行使得个人可以很方便地发表自己的意见和感想，对比商业和大众传媒，博客的信息具有更大的灵活性和真实性，因而会更多地影响个人决策。一些商业公司已认识到博客的重要性，把博客作为一种把握消费者行为动向、密切与消费者沟通的工具。

对网络结构和时序分析的方法可以合并用来进行基于博客信息的消费者行为分析。通过数据抽取后，研究人员利用可视化工具分析博客之间的链接，该工具可以可视化任意时间结点的链接结构，还可以通过滑动条来直观地调整视图，因此可以很容易地重现链接关系的时空演变。

研究人员认为博客内部的链接暗示着主题的传播，因此可以通过观察口述信息在博客间的传播来研究思想和行为的传播。如图 5 所示，一本通过博客传播变得非常流行的书，其原始博客，随着时间的推移，被越来越多的博客链接。

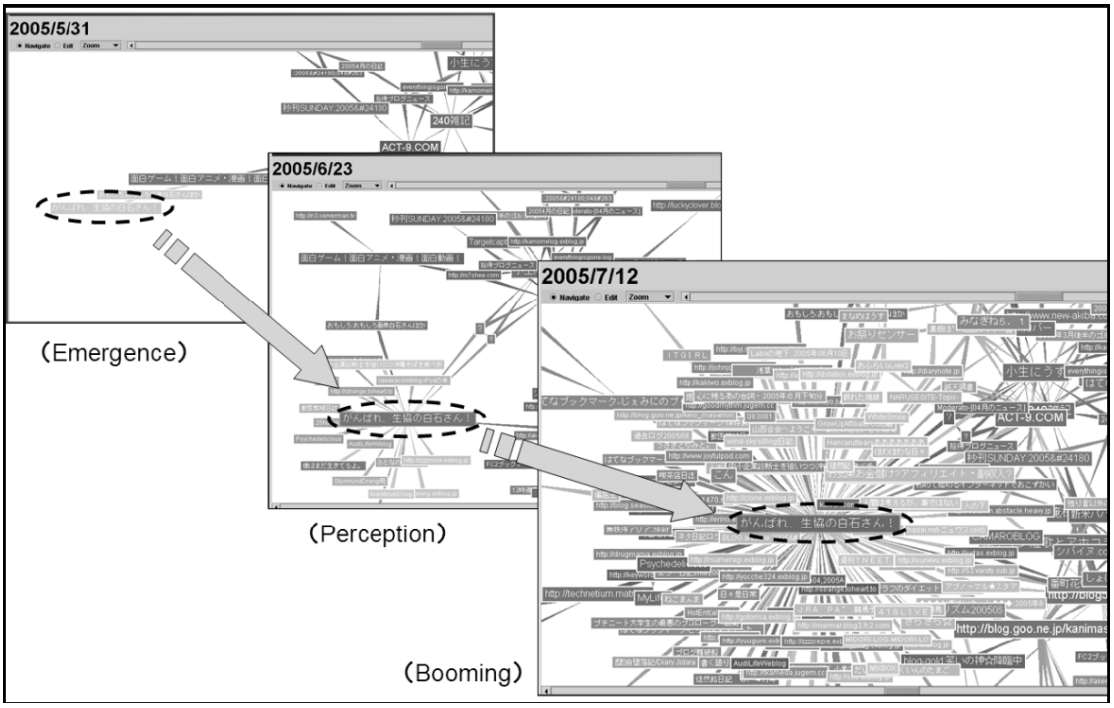

图 5 原始博客的链接随时间发生的变化[20]

可以通过同样的办法获得企业和他们的产品在同行业网站上的链接，这样就可以很容易地识别一些有吸引力的企业和品牌。

4 结语

Web archive 决不仅仅是 Web 页的历史集合，它更是一笔巨大的资源财富，拥有丰富的信息有待开发。Web archive 应用领域的每一个发展都为我们开辟了新的视野。

一些已经发展多年的技术，例如信息过滤、信息提取、文本挖掘、文本综述等，在某些特定领域的场合（例如电子商务）已经进行了成功的应用，但面对纷杂的、海量的网络信息，它们在效果和效率上都还有很大的改进空间。另外，海量信息的表现（呈现）也是一个引人入胜的领域，信息的类型、强度、相互之间的关系等时空特性，往往只有通过适当的表现才有可能带来有效的理解。因此数据的可视化也会为海量Web archive信息的应用带来精彩表现[21]。

Web archive 资源的应用已逐渐成为倍受关注的研究方向。开发合适的数据分析和挖掘工具是未来 Web archive 应用进一步拓展的关键。另外随着 Web 技术环境的变化，各种新出现的内容数据形式都将可能成为 Web archive 的新资源，同时也会对不同数据源的融合和挖掘技术提出新的挑战。

参考文献：

1 Internet Archive. http://www.archive.org/index.php. [2007-12-16]

[2] 中国Web信息博物馆，http://www.infomall.cn/ (2008-11-2)

3 Wayback machine，http://www.archive.org/index.php （2008-11-2）

4 Report on the 8th International Workshop on Web Archiving - IWAW 2008，http://www.dlib.org/dlib/november08/rauber/11rauber.html （2008-11-2）

[5] WEAR，http://archive-access.sourceforge.net/projects/wera/（2008-11-2）

6 XTF，http://www.cdlib.org/inside/projects/xtf/，（2008-11-2）

[7] Xinq, http://www.nla.gov.au/xinq/ （2008-11-2）

[8] Warrick. http://warrick.cs.odu.edu/，（2008-5-28）.

9 Lazy Preservation: Reconstructing Websites by Crawling the Crawlers,http://www.cs.odu.edu/~fmccown/pubs/lazyp-widm06.pdf (2008-11-2)

[10] Web Continuity, http://www.nationalarchives.gov.uk/webcontinuity/ (2008-11-8)

[11] 阎宏飞，可扩展 Web 信息搜集系统的设计、实现与应用初探,北京大学，博士论文，2002

[12] Rauber A., et al. Austrian Online Archive Processing: Analyzing Archives of the World Wide Web[J]. Research and advanced technology for digital libraries: 6th European conference, ECDL, 2002:16-31.

[13] Rauber A., et al. Uncovering Information Hidden in Web Archives[J]. D-Lib Magazine, 2002,8(12):1082-9873.

[14] Arms W.Y., et al. Building a research library for the history of the web[J]. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 2006:95-102.

[15] Arms W.Y., et al. A Research Library Based on the Historical Collections of the Internet Archive[J]. D-Lib Magazine, 2006,12(2):1082-9873

[16] Arms W.Y., et al. A Research Library Based on the Historical Collections of the Internet Archive[J]. D-Lib Magazine, 2006,12(2):1082-9873

[17] Masaru Kitsuregawa, Takayuki Tamura, Masashi Toyoda and Nobuhiro Kaji，Socio-Sense: A System for Analysing the Societal Behavior from Long Term Web Archive，Progress in WWW Research and Development，Springer Berlin / Heidelberg，2008

[18] 同 16

[19] 同 16

[20] 同 16

[21] 让社会科学插上信息技术的翅膀，http://cess.grids.cn/ourpdfs/Let%20social%20science%20ride%20on%20IT%20bullet%20train.pdf (2008-11-2)