提取模式和关系

从万维网

谢尔盖·布林

计算机科学系

斯坦福大学

sergey@cs.stanford.edu

抽象。万维网是一个巨大的信息资源。同时，它是极其分散。特定类型的数据（如餐厅列表）可能以多种格式分散在数千个独立的信息源中。在本文中，我们考虑了提取此类数据类型fr的关系的问题。所有这些源都自动提取。我们提出了一种利用模式集和关系之间的二元性的技术，从小样本开始发展目标关系。为了测试我们的技术，我们用它来从万维网上提取（作者，标题）对的关系。

* 介绍

万维网提供了几乎所有类型的信息来源，从DNA数据库到简历，再到最喜欢的餐馆列表。但是，此信息通常分散在许多 Web服务器和主机中，使用许多二次格式。如果可以从万维网中提取这些信息块并集成到结构化形式中，它们就会形成前所未有的信息来源。它将包括最大的international人名录，最大和最多样化的产品数据库，最大的学术著作书目，和许多其他有用的资源。

在使用speci盟友编码包装器或lters[Tsi，MOS97]集成一些信息源方面，已经做了大量工作。但是，创建这些源可能非常耗时，通常用于数十个源，而不是数千个源。在本文中，我们解决了从万维网上可能持有关系饼的数千个源中提取关系的问题。我们的目标是发现信息来源，并完全自动地从信息源中提取相关的形成，或者以极低的人工干预。

在本文中，我们考虑了从Web中分离出书籍关系的问题。（作者、标题）对。直观地讲，我们的解决方案工作原理如下。我们从一小组种子（作者，标题）对开始（在测试中我们使用了一组只是ves的账本）。然后，我们在网上出现这些书。从这些事件中，我们认识到书籍引文的模式。然后，我们在网上搜索这些模式和新书。然后，我们可以拿这些书和

和它们的所有发生，并从生成更多的模式。我们可以使用这些新模式来出版更多的书籍，等等。最终，我们将获得大量书籍和模式，用于出版它们。

* 模式与关系的二元性

我们提出的方法称为DIPRE-双迭代模式关系前泛。它依赖于模式和关系之间的二元性，我们在下面解释。

2.1 问题

在这里，我们将更正式地解决我们的问题：

让 D 成为非结构化信息的大型数据库，如万维网。让 R = r1;：：; rn是目标关系。R 的每个元组、t、R 在 D 中出现一次或多次。每个此类事件都由 t 的所有 eld 组成，表示为字符串，在 D 中彼此非常接近（在 Web 中，这意味着所有 eld 在同一 Web 页上彼此接近）。

在本文中检查的测试问题中，目标关系 R 是 Web 上发生的一组书籍 （作者、标题） 对。显然，这还不好。然而，鉴于潜在的作者和标题，以及它们在网络上被提及的地方，人类通常可以判断这是否是合法的

书。 0

如果我们计算一个近似值，R 的 R0，则覆盖范围为 jRj  =jRj 和

R

错误率为jR0  ，Rj .我们的目标是最大化覆盖范围和最小化错误

jR0 j

率。但是，低错误率比高覆盖率重要得多。给定一个大型数据库 D，仅召回 20% 是可以接受的。但是，错误率超过 10%对于许多应用程序来说可能毫无用处。

通常，我们无法实际计算 R。因此，我们不能不知道覆盖率和错误率的精确值。但是，我们可以通过让用户检查 R0的随机元素来对错误率进行采样。报道是更多的邪教估计。

2.2 模式

直观地，模式与目标关系的元数的一种特定格式匹配。理想情况下，模式是规格 c 不够匹配任何元数，不应该在相对关系n， 但是， 在实践中可能会出现一些误报。模式可能具有各种表示形式。在我们的工作中，我们使用的正则表达式类非常有限。更正式：

让 p 成为一种模式。然后 MD（p） 是匹配 D 中的 p 的元组，jpjD是 MD（p）中的元素数。然后是p，CD的覆盖范围（p;R） = jMD（p） = Rj\_jRj，p 的错误率为 ED（p;R） = jMD（p） ， Rj\_jMD（p） j.

对于一组模式，P = p1;：：：;pk，我们 de ne MD （P） = Sp2P MD（p）。我们扩展 CD （P;R）和E D（P;R） 类似。MD （P ） 的替代设计可能需要一个元组来匹配多个模式（参见第 6 节）。

2.3 模式关系二元性

一个重要的观察是，给定一组模式，P具有高覆盖率和低错误率，我们可以通过将所有匹配与所有模式进行匹配来构造一个很好的 R 近似值。因此，给定一组良好的模式，我们可以构建一组良好的元组。然而，我们也希望有相反的属性 - 给定一组良好的 tuples，我们可以构建一组良好的模式。我们可以通过在 D 中找出元数的所有匹配项并发现事件中的相似性来执行此操作。从元数中分部和模式组合的能力为我们提供了强大的力量，这是我们在本文中提出的技术的基础。

* 双迭代模式关系提取

双迭代模式关系提取- DIPRE是利用模式关系二元性提取关系的技术。其工作原理如下：

1. R0 示例

从目标关系的 R0小样本开始。此示例由用户提供，可能非常小。在我们的测试中，我们使用了作者的一本书列表。

1. 查找事件 （R0;D）

然后，以及 D 中 R0的元数的所有匹配项。在我们的实验中，这些是作者的附近事件和文本书籍的标题。与找到的元组一起，保留每个事件的上下文（url 和周围的文本）。

1. P 一代模式 （O）

根据事件集生成模式。这是算法中棘手的部分。粗略地说，此例程必须为具有类似上下文的事件集生成模式。模式需要有一个低

错误率，因此重要的是它们不要过于笼统。图案的覆盖率越高越好。但是，低覆盖率可以通过更大的数据库进行补偿。

4. R0 MD（P ）在数据库中搜索匹配任何模式的元数。

5. 如果 R0 足够大，则返回。否则转到步骤 2。

3.1 控制扩展

上述过程不一定非常稳定，并可能偏离R。特别是，在MD （P） 的几个假元组成可能导致几个假模式

* 在下一次迭代中。这反过来又会导致一大堆假元。对于

因此，GenPatterns 例程必须小心，以尽量减少潜在假元组（或几个小元组）造成的损坏量。另一种安全措施是更加严格地去MD（P），以便要求元数与P中的多种模式相匹配。在我们执行的测试中，第二个度量值是不必要的，但在 future 测试中可能是必要的。最后，随着关系的扩大，各种关系可能需要进行统一。

* 查找作者和标题

对于我们的实验，我们选择从万维网上计算（作者、标题）对的关系。这个问题对DIPRE特别适合，因为有许多知名书籍列在许多网站上。许多网站都遵循整个网站的合理统一格式。

4.1 书籍模式

为了使用DIPRE到书籍，它是必要的，以去什么模式组成。模式的确定在很大程度上决定了DIPRE的成功。但是，对于我们的测试，我们使用的模式非常简单。它需要进一步调查，以确定是否更复杂的拍子工作得更好。

我们将一个模式作为一个 ve-元组：（顺序、urlpre x、前 x、中间、su x），其中顺序是布尔值，其他属性是字符串。如果顺序为 true，则（作者、标题）对与集合（WWW） 中的d ocment 与 URL 匹配URL\*匹配的模式匹配，并且哪个文本与正则表达式：

[前缀、作者、中间、标题、后缀]

作者仅限于：

[A-Z][A-Za-Z.，&]5 .;30[A-Za-Z.]

标题仅限于：

[A-Z0-9][A-Za-z0-9.，'&]4;45[A-Za-z0-9？！

如果顺序为 false，则切换标题和作者。

4.2 发生

我们还必须了解事件的结构，因为它应该与模式的 de nition 相对应。 （作者、标题）对的出现由七组组成：

（作者，标题，顺序，URL，前x，中间， 苏x）

顺序对应于标题和作者在文本中发生的顺序。URL 是它们出现红色的文档的URL。前 x 由作者之前的 m 字符（在测试中 m 为 10）组成（如果标题为

rst）。中间是作者和标题之间的文本，su x 由标题（或作者）之后的 m 字符组成。1

4.3 为书籍生成模式

DIPRE程序的一个重要组成部分是GenPatterns例程，它获取一组书籍，并将其转换为模式列表。这是一个不平凡的问题，有整个模式识别致力于解决这个问题的一般版本。但是，为了我们的目的，我们使用一组简单的启发式方法从事件生成模式。只要很少误报（生成非书籍的模式），这就是科学。每种模式只需要非常小的覆盖范围，因为网络是广阔的，有许多信息来源，所以所有模式的总覆盖率仍然可以是实质性的。

假设我们得到了一组事件，我们希望构造一个可能匹配所有事件的规范模式。我们可以按照以下方式执行此操作：

1. 验证所有事件的顺序和中间数是否相同。如果不是，则不可能生成一个模式来匹配它们。设置模式.顺序和出型模式.中间到订单和中间分别。
2. 查找所有 URL 中匹配时间最长的前 x。设置模式.urlpre x 到该前 x。
3. 将outpattern.pre x设置为发生前 x的最长匹配 su x。
4. 将outpattern.su x设置为匹配时间最长的 su x的匹配前x 的匹配。

我们表示这个常规的GenOne模式（O）。

模式规格城市类似上述生成的模式可能过于笼统或过于规范。我们不关心它太特别c，因为将有许多模式生成和结合将有许多书籍。然而，模式可能太基因ral，并可能产生许多非书籍。

为了解决这个问题，我们试图测量模式的规格城市。模式p的指定城市大致对应于，但（X 2 M D（p））），其中X是一些随机变量，均匀分布在R的元数域上。 2为了快速计算，我们使用以下公式斯佩奇市模式（jsj 表示 s 的长度）：

规格城市（p） = jp：中间jjp：urlpre xjjp：前xjjp：su xj

* 如果行结束或开始接近发生，则前 x 和 su x 实际上可能小于m字符，但这是当前实现的限制

也不清楚它是否有正或负等。

* 如果域与所有字符串的空间一样，则统一分布可能不太合理，应使用 di erent 分布。

我们拒绝任何模式与过低的规格城市，以便不生成过于一般性的帕特特。更规范的，我们坚持，规格城市（p）n > t n 是支持模式 p 和 t 的匹配项的匹配数的书籍数量。这可确保模式的所有字符串都非空（从其他情况而言，指定城市为零）。此外，我们要求基于一个示例的模式的n > 1 ince 非常容易出错。

生成模式的算法在这里，我们提出了一代模式（O）的算法。它利用了第 4.3 节中引入的算法 GenOnePattern（O）。

1. 按顺序和中间将 O 中的所有事件分组。让生成的组为 O1;：：Ok。
2. 对于每个组 Oi， p GenOne 模式 （Oi）。如果 p 满足特定城市重新要求，则输出 p.否则：

\*如果所有O在O我具有相同的URL，然后拒绝O i。

\*否则，将 Oi 中的出现o 分隔到子组中，这些子组中的字符是过去 p.urlpre x 的一个。

当生成的模式不是 su 科学规范 c 时，此例程使用基于 url 的简单进一步细分。也可以想象使用前 x 或 su x。

我们描述了一种从事件清单中生成模式的简单技术。人们可以想象更先进的技术，这是进一步研究的主题。然而，正如结果（第5节）所示，即使是这个简单的方案也运作良好。

4.4 性能问题

有两个非常苛刻的任务DIPRE - 给定一长串的书籍和列有模式匹配的图书的匹配。这两个操作都必须在非常大的 Web 文档数据库上执行。

对于 rst 任务，即书籍的出现，我们不可能通过两个 fgrep lters 传递数据。一个只穿过包含有效作者的行，另一个只穿过包含有效标题的行。在此之后，使用 Python 编写的程序的任务是实际检查行中是否有匹配的作者和标题，标识它们并生成s 输出。为此，我们尝试了几种替代方法，涉及 Flex 和 Python 中的大型正则表达式，但它们很快超过了各种内部边界。

对于第二个任务，我们只使用 Python 程序。每个模式都转换为一对正则表达式，一个用于 URL，一个用于实际发生。每个 URL 都经过 rst 测试，以查看哪些模式适用于它。然后，程序测试相关正则表达式的每一行。这种方法相当缓慢，需要改进。未来版本可能会使用 Flex 或

rex C 库。通过仅定位与模式匹配的 URL，我们尝试这样做会稍微简化此任务。如何，数据的结构并不是让这一点变得微不足道，我们希望我们开发的技术足够通用，能够处理 URL 上没有任何限制。

此时，从事件生成模式并不是一个性能问题，因为只有数千个事件是 geneese。随着运行更大的测试，这将变得更加重要。目前，事件是按顺序和中间排序的。然后，Python pro-gram 读取结果列表并生成模式，如第 4.3 节所述。

* 实验

虽然迄今进行的实验非常有限，但由于时间限制，它们产生了非常积极的结果。更多的实验正在进行中。

5.1 实验中使用的 Web 数据

对于数据，我们使用了 2400 万个网页的存储库，总计 147 GB。这些数据是斯坦福网站基础的一部分，用于谷歌系统引擎[BP]和其他研究项目。作为搜索引擎的一部分，我们构建了整个存储库的倒转索引。

存储库跨越许多磁盘和多台计算机。即使不进行任何实质性处理，也只需一次传递数据，也需要大量的时间。因此，在这些迭代中，我们仅在任何给定迭代上对存储库的子集进行传递。

这个项目的一个重要方面是，存储库几乎没有来自亚马逊[Ama]的网页。这是因为他们自动生成的 url 使爬网的邪教。

5.2 模式关系扩展

|  |  |
| --- | --- |
| 艾萨克·阿西莫夫 | 黎明的机器人 |
| 大卫·布林3 | 星潮上升 |
| 詹姆斯·格里克 | 混沌：创造新科学 |
| 查尔斯·狄更斯 | 伟大的期望 |
| 莎士比亚 | 错误的喜剧 |

图1.书籍的初始样本。

|  |  |
| --- | --- |
| URL 模式 | 文本模式 |
| www.sff.net/locus/c. | <LI><B>标题</B> 按作者（ |

~

dns.city-net.com/lmann/awards/hugos/1984.html <i>标题</i>

~

dolphin.upenn.edu/dcummins/texts/sf-award.htm作者 |标题

作者（

||(

图2.在 rst 迭代。

我们只用5本书开始了实验（见图1）。这些生成了 199 个事件并生成了 3 个模式（参见图 2）。有趣的是，只有两本的电子书产生了这些图案，因为它们都是科幻书。通过匹配 URL 的 pro 运行这些模式，可引入 4047 唯一（作者、标题）对。他们大多是科学学的，但也有一些例外。（参见图 3。

|  |  |
| --- | --- |
| H. D. 埃弗雷特 | 死亡面具和其他幽灵 |
| H. G. 威尔斯 | 月球上的第一个男人 |
| H. G. 威尔斯 | 科幻小说：第2卷 |
| H. G.威尔斯 | 月球上的第一个男人 |
| H. G. 威尔斯 | 隐形人 |
| H. G. 威尔斯 | 莫罗博士岛 |
| H. G. 威尔斯 | 科幻小说第1卷 |
| H. G. 威尔斯 | 事物形态：终极革命 |
| H. G. 威尔斯 | 时间机器 |
| H. G. 威尔斯 | 世界战争 |
| H. G. 威尔斯 | 当睡眠者唤醒 |
| H. M. 胡佛 | 穿越空旷的旅程 |
| H. P. 洛夫克拉夫特和八月德莱思 | 门槛处的卢克尔 |
| H. P. 洛夫克拉夫特 | 在疯狂山脉和其他恐怖故事 |
| H. P. 洛夫克拉夫特 | 查尔斯·德克斯特·沃德案 |
| H. P. 洛夫克拉夫特 | 来到萨纳特和其他故事的厄运 |

图3.在 rst 迭代。

在搜索了大约500万个网页时，发现了3972本这些书。这个数字令人失望，因为它不像在 rst 迭代中那样大爆炸。但是，至少需要几天时间才能对整个存储库运行，因此我们没有尝试生成更多。这些事件产生了 105 个模式，其中 24 个模式具有 url 前 xes，这些模式不是完整的 url。一次通过两百万个网址产生了9369个独特的（作者，标题）对。不幸的是，其中有一些假书。特别是，其中242个是正当的书名，但有一个作者的[结论]。我们从列表中删除了这些。这是整个过程中唯一的人工干预。在未来的实验中，它会

有趣的是，看看离开这些是否会产生额外的垃圾量。

对于 nal 迭代，我们选择使用存储库的子集，该子集对工作簿进行协调。其中包括大约156，000份文件。扫描剩余的9127本书产生了9938次。这些反过来又使346模式。扫描同一组文件产生了15257本独特的书籍，几乎没有虚假数据。（参见图 4）

这个实验正在进行中，希望一个更大的书籍列表将很快被基因化。当前一个是在线[Bri] 。

5.3 结果质量

为了分析结果的质量，我们从列表中挑选了20本随机书籍，并试图通过在亚马逊[Ama]、Visa购物指南、斯坦福在线图书馆目录和网络上搜索来验证它们是否真实。4作为衡量结果质量的一个指标，20个中，有19个都是书本。剩下的书实际上是一篇文章 - [为什么我投票给用户汽车"，由安德鲁托比亚斯。

令人大吃一惊的是，除了网络之外，有些或所有来源都找不到这些书籍。这些书有些是网上书;有些是晦涩或绝版;有些只是没有列出在一些网站没有明显的原因。总共，20本书中有5本不在亚马逊上，亚马逊声称有250万册图书的目录。

除了上面提到的文章之外，数据还有一些明显的问题。有些书被提到几次，由于小的，如资本化，间距，作者是如何上市（例如[E.R. Burroughs'vs \_Edgar Rice Burroughs"）。然而，幸运的是，作者对于他们的名字是如何列出的非常特殊的，这些重复是有限的。在一些情况下，一些信息被追加到 author 的名称上，如发布日期。

* 结论

我们的总体目标是能够利用整个万维网的庞大性，从整个万维网中提取结构化数据。DIPRE 已被证明是一个了不起的工具，在简单的例子，即列出书籍列表。它从一套5本书的样本开始，并扩展到一个相对高质量的清单，超过15，000本书，很少人为干预。同一工具可能应用于许多其他领域，如电影、音乐、餐馆等。 此工具的更复杂的版本可能能够提取人员目录、产品目录等。

* 不幸的是，国会图书馆的搜索系统在这些测试时已经关闭。

|  |  |
| --- | --- |
| 亨利·詹姆斯 | 欧洲人 |
| 亨利·詹姆斯 | 金碗 |
| 亨利·詹姆斯 | 女士的肖像 |
| 亨利·詹姆斯 | 螺钉的转动 |
| 亨利·詹姆斯 | 螺钉转动 |
| 亨利·约翰·可乐 | 滚石的踪迹 |
| 亨利·罗 | 基督教历史上的地标 |
| 亨利·基索 | 西风 |
| 亨利·劳森 | 在世界广阔的时代 |
| 亨利·朗费洛 | 海华沙之歌 |
| 亨利·米勒 | 热带癌症 |
| 亨利·彼得罗斯基 | 设计发明 |
| 亨利·彼得罗斯基 | 有用事物的演变 |
| 亨利·罗斯 | 称之为睡眠 |
| 亨利·萨姆纳·缅因 | 古代法律 |
| 亨利·塔克曼，林赛，菲拉 | 文学的特点 |
| 亨利·范·戴克 | 蓝花 |
| 亨利·范·戴克，斯里布 | 第 O 天 |
| 亨利·韦尔马 | 皮特·斯图维森特的生活与时代 |
| 亨利·沃兹沃斯·朗费洛 | 保罗·里维尔的骑 |
| 亨利·沃兹沃斯·朗费洛 | 伊万杰林 |
| 亨利·沃兹沃斯·朗费洛 | 海华沙之歌 |
| 赫伯特·唐纳德 | 林肯 |
| 赫伯特·哈特 | 西北老堡 |
| 赫伯特·梅森 | 拉斐特埃斯卡德里耶 |
| 赫伯特·洛特曼 | 儒勒·凡尔纳：探索传记 |
| 赫伯特·斯宾塞 | 人与国家 |
| 赫尔曼·戴利 | 为了共同利益 |
| 赫尔曼·戴利 | 珍惜地球 |
| 赫尔曼·基特里奇 | 英格索：传记欣赏 |
| 赫尔曼·哈肯 | 大脑功能原理 |
| 赫尔曼·黑塞 | 德米安 |
| 赫尔曼·黑塞 | 悉 |
| 赫尔曼·黑塞 | 西达塔 |
| 赫尔曼·梅尔维尔 | 巴特莱比，斯克里维纳 |
| 赫尔曼·梅尔维尔 | 比利·巴德 |
| 赫尔曼·梅尔维尔 | 比利·巴德 |
| 赫尔曼·梅尔维尔 | 鲸 |
| 赫尔曼·梅尔维尔 | 康恩斯人 |
| 赫尔曼·梅尔维尔 | 恩坎塔达岛，或魔法岛 |
| 赫尔曼·梅尔维尔 | 类型：波利尼西亚生活的窥视 |
| 赫尔曼·韦斯 | 日落侦探 |
| 赫尔曼·沃克 | 战争与纪念 |
| 赫尔曼·黑塞 | 克林索的《最后的夏天》 |
| 赫尔曼·黑塞 | 克努普 |
| 赫尔曼·黑塞 | 罗沙尔德 |
| 赫尔曼·黑塞 | 来自另一颗星星的奇怪消息 |
| 希罗多德 | 历史 |
| 希罗多德 | 历史 |
| 希罗多德 | 希罗多德的历史 |
| 赫歇尔·霍布斯 | 牧师手册 |
| 赫歇尔 | 第一阶段：月亮 |
| 希亚森 | 暴风雨 |
| 希莱尔 | 竞争对手和新到达 |
| 希莱尔 | 大异一代 |
| 希拉里·贝利 | 卡桑德拉：特洛伊公主 |
| 希拉里·诺曼 | 苏珊娜的钥匙 |
| 希尔伯特·申克 | 计时序列 |
| 希尔伯特·申克 | 阿巴科礁之战 |
| 希尔达·康克林 | 小女孩的诗 |
| 希尔达·休斯 | 颤动 |
| 希尔达·休斯 | 当教会院家雅恩 |
| 希勒曼 | 时间的窃贼 |
| 希勒曼 | 皮肤步行者 |
| 希勒曼 | 会说话的上帝 |
| 希拉姆·科森 | 布朗宁简介 |
| 哈马尔·霍思·博耶森 | 挪威的童年 |
| 哈马尔·霍思·博伊森 | 来自两个半球的故事 |
|  |  |

图4。书籍样本 nal列表。

6.1 可扩展性和稳定状态

此方法的可伸缩性存在若干挑战。一个是扫描大型存储库上大量模式和元数所需的每形式。基础算法和实现的改进可能在不久的将来解决这个问题。

一个潜在的更邪教的障碍是，DIPRE是否可以保持偏离目标，因为它扩大关系。例如，由于它实际上只使用了种子样本中的两本科学词汇，为什么它不产生大量的科学词汇表。显然，它被吸引到所有书籍的汇编，甚至一些散播的文章设法进入关系。随着关系的扩大，控制这一点不平凡，但有几种可能性。

连接到奇异值分解一种可能性是重新填充 MD（P ） 以需要多个模式来匹配元组。更极端的版本是为每个元组和图案分配权重。根据匹配的图案的权重，为匹配的元组分配权重。生成的模式根据与之匹配的元数的权重分配权重。如果以线性方式完成此操作，则此技术分解为元组模式矩阵的奇异值分解（乘以其转置）。这类似于在文档词矩阵上完成的延迟语义索引 [DDF=90]。在这种情况下，最终的稳定状态是占主导地位的igenvector。不幸的是，这与初始样本无关，这显然不可取。尽管如此，与LSI的关系是令人信服的，并需要进一步的投资。

即使没有使用权重，稳定状态与上述初始状态的独立性也可能是一个问题。有几种可能的解决方案。其一是只运行有限的迭代次数，如本文中演示的那样。另一种解决方案是确保元数转换为元数是非线性的，并且具有一些依赖于初始状态的局部稳定状态。这可以通过使用初始

样本 R0在 GenP atterns的计算中。在这种情况下，用户也可能为

提供 R ， 反例列表.

6.2 自动提取的影响

这个实验最令人惊讶的结果之一是把没有列在主要在线来源的书籍，如道格拉斯·克拉克[Cla]在网上出版的《被解散》一书，或者多莉·拉德福德[Rad04]的《年轻的园丁卡伦达尔》一书。1904年出版的模糊作品。如果可以展开图书列表，并且联机源中列出的几乎所有书籍都可以进行拉伸操作，则生成的列表可能比任何现有图书数据库都更完整。生成的列表将是数千个小型在线来源的产品，如

与目前的书籍数据库相反，这些数据库是几个大型的出版来源的产物。这种信息的变化可以有重要的社会信息。

引用

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| [爱] | 亚马逊主页。http://www.amazon.com. | | | | |  |
| [BP] | 谢尔盖·布林和 | | 拉里 | 网页。 | 谷歌搜索引擎。 | http://google。 |
|  | stanford.edu. |  |  |  |  |  |
| [布里] | 我是谢尔盖·布林 | 列表 | | 书。 | http://www-db.stanford.edu/~sergey/ | |
|  | 书单.html. |  |  |  |  |  |
| [克拉] | 道格拉斯·克拉克已解散.本杰明出版社， 69 希尔克雷斯特驱动器， 巴斯 Ba2 | | | | | |
|  | 1HD，英国。http://www.bath.ac.uk/~exxdgdc/poetry/library/di1.html. | | | | | |
| [DDF=90] | 斯科特·迪尔韦斯特、苏珊·杜马斯、戈尔格·富纳斯、托马斯·兰道尔和 | | | | | |
|  | 理查德·哈什曼通过潜在语义分析编制索引。 日记本 | | | | | |
|  | 美国信息科学学会， 41（6）：391\_407， 1990. | | | | | |
| [莫斯97] | 半结构化数据管理研讨会。http://www.research。 | | | | | |
|  | att.com/~suciu/workshop-papers.html， 1997年5月. | | | | |  |
| [Rad04] | 多莉·拉德福德 | | 的 | 年轻 | 园丁卡伦达尔。 | 亚历山大·莫尔 |
|  | 1904年，伦敦。http://www.indiana.edu/~letrs/vwwp/radford/ | | | | | |
|  | 日历. html. |  |  |  |  |  |
| [Tsi] | 齐姆米斯家 | 网页。 | | http://www-db.stanford.edu/tsimmis/tsimmis。 | | |
|  | html. |  |  |  |  |  |
| [Vis] | 签证购物指南书籍。http://shopguide.yahoo.com/shopguide/ | | | | | |
|  | 书籍.html. |  |  |  |  |  |