主题 Web 信息采集的研究与设计*

李盛韬 吴丽辉 于满泉 潘文锋 余智华 王斌 程学旗

中国科学院计算技术研究所 软件研究室 北京 100080
E-mail: lishengtao@software.ict.ac.cn

摘 要：主题 Web 信息采集是信息检索领域内一个将采集技术与过滤方法结合的新兴方向，也是信息处理技术中的一个研究热点。本文分析了主题 Web 信息采集的基本问题，提出了难点以及相关的解决方案，并在此基础上设计了“天达”主题 Web 信息采集系统。

关键词： 信息采集；信息检索；信息处理；主题

Research and Design of Focused Web Crawler

Li Shengtao Wu Lihui Yu Manquan Pan Wenfeng Yu Zhihua
Wang Bin Cheng Xueqi

Software Division, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080
E-mail: lishengtao@software.ict.ac.cn

Abstract: Focused web crawling is a new crawling direction in the field of information retrieval which is combined with filtering methods. And it also is a research hotspot in the information processing technologies. This paper argues the principles, difficulties and measures of the focused web crawler, and then detailedly analyses the design of our SkyReach focused web crawler.

Keywords: Web Crawler; Information Retrieval; Information Processing; Focused Crawler

1 引言

随着 Internet 的迅速发展，网络正深刻地改变着我们的生活。它在给人们提供丰富信息的同时，又给人们提出巨大挑战。因此，Web 信息采集、发布和相关处理日益成为人们关注的焦点。

传统的 Web 信息采集的目标就是尽可能多地采集信息页面，而较少考虑采集页面的准确性，它存在着很多缺陷。随着 WWW 的爆炸性增长，信息采集的速度越来越不能满足实际需要。最近的试验表明，即使大型的信息采集系统，对 Web 的覆盖率也只有 30-40%。

主题采集则可以通过对整个 Web 按主题分块采集，并将不同块整合，来提高整个 Web 的采集覆盖率。对于传统的信息采集来说，刷新一遍需要数周到一个月的时间[1][2]，这使得页面的失效率非常巨大。一个好的缓解办法就是采用主题采集，通过减小采集页面的数量来

*本课题受国家 973 项目（G1998030413）和中科院计算所领域前沿青年基金（20016280-8）资助。
李盛韬(1976-)，男，甘肃兰州人，硕士，主要研究方向：信息检索，数据挖掘，分布式系统。

减小刷新时间，进而减小已采集页面的失效率。传统的信息采集需要消耗很多系统和网络资源，而它们中大部分利用率很低，基于主题的采集有效地提高了采集到页面的利用效率。

本文内容组织如下：第二章介绍相关的研究；第三章“天达”主题 Web 信息采集系统的结构和实现情况；第四章是实验结果；最后一章展望了主题 Web 信息采集的发展动向。

2 相关研究

基于主题的 Web 信息采集(Focused Crawling)，也称为 Topic-Specific Crawling，主要是指选择性地搜寻那些与预先定义好的主题集相关的页面进行采集的行为。它是 Web 采集中最重要的一种类型。

Kleinberg[7]使用 Authorities/Hubs 算法实现了对广泛主题页面的采集，但对具体主题效果不佳。Cho[4]描述了一种 focused crawling 技术，即首先采集那些对主题更重要的页面，过滤方法可以用相似度、Backlinks、PageRank 以及 Location 等。Chakrabarti[3]则在他们的 Focused Crawling 系统中利用主题信息的 Linkage Locality 特性(页面趋向于拥有链接到它的页面的页面主题)和 Sibling Locality 特性(对于链接到某主题页面的页面，它所链接到的其它页面也趋向于拥有这个主题)来进行过滤算法的设计。Diligenti[5]改进了 Linkage/Sibling Locality 特性，并建立了一张层式的上下文图。

Hersovici[6]的主题采集器用向量空间模型方法计算采集到页面与主题词之间的相似度，并将此值在衰减因子的处理下传递给它所链接的页面，然后结合链接周围的文本决定是否采集此链接。Aggarwal[1]的 Intelligent Crawling 系统则利用 Content、URL Token、Link 和 Sibling 进行主题预测，该方法的好处是通过少量关键词就可以开始搜索，无需事先对主题页面分类。

McCallum[9]利用 Naive Bayes 方法来选择待采集链接，他首先根据文本和链接的扩展元数据对链接进行分类，通过训练和学习的结果来判别每个链接的得分，试验结果显示，它的优势是能快速发现更多的学术论文。

Menczer 则评价了三种基于主题采集的策略:1).Best first Crawler(通过计算链接所在页面与主题的相似度来得到采集优先级)；2).PageRank(通过每 25 页计算一遍 PageRank 值来得到采集优先级)；3).InfoSpiders（通过链接周围的文字，利用神经网络和遗传算法来得到采集优先级）。经过试验比较，作者发现，Bestfirst 方法最好，InfoSpiders 方法次之，PageRank 算法最差。

从上面的说明，我们可以看出当前的研究主要集中在利用链接分析、扩展元数据、向量空间模型、Naive Bayes、神经网络和遗传算法等方法来进行主题过滤。

整个 Web 上的页面主题分布是混杂的，但同一个主题在 Web 上分布却有一些规律。我们将之总结为四个特性：Hub 特性、Sibling/Linkage Locality 特性、站点主题特性、Tunnel 特性。

康奈尔大学的教授 Jon M. Kleinberg 发现 Web 上存在大量的 Hub 页面，这种页面不但含有许多 outlink 链接，并且这些链接趋向于相关同一个主题。也就是说，Hub 页面是指向相关主题页面的一个中心。我们把主题在 Web 上的这一特性称为 Hub 特性。

在 Hub 特性的基础上，人们又提出了 Sibling/Linkage Locality 特性[1]。1).Linkage Locality,即页面趋向于拥有链接到它的页面的主题；2).Sibling Locality,对于链接到某主题页面的页面，它所链接到的其它页面也趋向于拥有这个主题。这实际上是 Hub 特性的变形，主要是从设计者设计的角度考虑的。我们称之为 Sibling/Linkage Locality 特性。

一个站点趋向于说明一个或几个主题，并且那些说明每个主题的页面较紧密地在此站点内部链接成团，这就是站点主题特性。我们认为，这主要与网站的设计思路有关。每个网站在设计时都有目标，而这种目标往往就集中在一个或几个主题中。而网站的浏览者往往也有

一定的目的性，这个目的性体现在用户趋向于浏览同一主题的页面。

在 Web 中还有一类现象，就是主题页面团之间往往需要经过较多的无关链接才能相互到达。这些无关链接就像长长的隧道，连接着两个主题团，我们把这种现象称为“隧道现象”，也称为 Tunnel 特性。在基于主题的页面采集过程中，Tunnel 的存在极大地影响着采集的质量，它是一个重要的难题。

以上四个特性存在着一定的关系。Hub 特性说明了主题容易成团出现的现象，Linkage/Sibling Locality 特性进一步对成团的特性有所扩展，站点主题特性说明了主题团所在的位置(即大部分分布于站点的内部)，而 Tunnel 特征说明了主题团在 Web 上的分布并不稠密。我们将根据这些规律，设计主题过滤算法。

3 基于主题的 Web 信息采集系统模型

3.1 系统模型

我们在国内外已有主题采集系统的基础上，设计了“天达”主题采集系统。为实现对基于主题的信息自动采集，我们将整个处理过程分成七大模块：主题选择、初始 URL 选择、Spider 采集、页面分析、URL 与主题的相关性判定(链接过滤/链接预测)、页面与主题的相关性判定(页面过滤)、数据存储。

3.2 主题的选择和采集起点的选择

为了有效地进行主题采集，需要考虑的一个重要问题就是主题选择。针对随便的主题词可能较大地影响采集效果，系统一般提供给用户一个主题分类目录以供选择。为了有效地确定用户选定主题的含义，用户要提供对主题的进一步描述，比如提供若干表达主题含义的文本。我们的系统是按中国图书馆分类方法的一级和二级目录对主题进行分类的，并在每个主题下配备了一些主题文本，以供用户选择。

采集器是从一个种子 URL 集出发，通过 Web 协议向所需的页面扩展的。根据 Linkage/Sibling Locality，系统需要选择质量较高的主题 URL 作为初始种子 URL 集。

3.3 **Spider** 采集

这个部分处于系统的底层，也叫“网络蜘蛛”，是系统专门与具体的 Web 打交道的部分。它主要通过各种 Web 协议来自动采集 Internet 上 WWW 站点内有效的信息（包括文本、超链接文本、图像、声音等各类文档）。目前系统实现的主要是针对 HTTP 协议的。这一部分的主要任务是将全局 URL 队列中的 URL 分配给各个 Spider 采集器，Spider 采集器的个数根据系统的需要动态分配，如图 1 所示。

3.4 页面分析

在页面采集到以后，我们要从中提取出链接、元数据、正文、标题、摘要来，以便进行后续的过滤和其它处理。我们在这里主要介绍链接和标题的提取。

3.4.1 链接的提取

对抓取到的页面需要分析其中的链接，并对链接中的 URL 进行必要的转换。首先判别页面类型，显然只有类型为“text/html”的页面才有必要分析链接。页面的类型可由应答头分析得出，有些 WWW 站点返回的应答信息格式不完整，此时须通过分析页面 URL 中的文件扩展名来判别页面类型。遇到带有链接的标记如<A>、<AREA>、<FRAME>等，就从标记结构的属性中找出目标 URL，并从成对的该标记之间抽取出正文作为该链接的说明文字(扩展元数据)。这两个数据就代表了该链接。

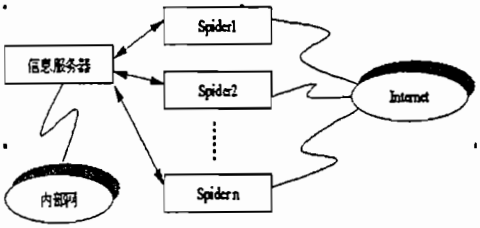

图 1

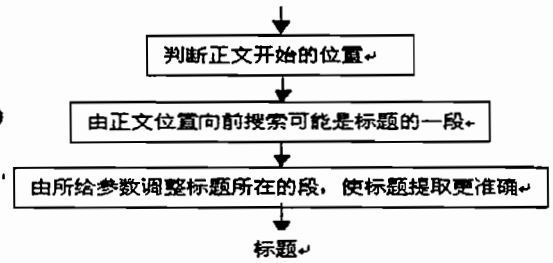

图 2

3.4.2 标题的提取

如图 2 所示，标题的提取分为三步：1).判断正文开始的位置，从文章开头开始，逐段扫描，直到某一段长度不小于设定的正文最小长度，就假定这段为正文中的一段。2). 由正文位置向前搜索可能是标题的一段，根据字体大小、是否居中、颜色变化等特征找出最符合的一段文字作为标题。3). 由所给参数调整标题所在的段，使标题提取更准确。句法、语义、统计分析标题段的前后几段，以准确确定标题段的真实位置；向前或向后调整几段，追加前一段或后一段。

3.5 URL 与主题的相关性判定

为了有效地提高基于主题的 Web 信息采集的可靠性(查全率和查准率的综合)和效率，系统需要在采集过程中增加过滤机制，以使得采集的页面能够向主题靠拢。过滤方法主要有四种：根据元数据的过滤、根据扩展元数据的过滤、根据链接分析的过滤、根据页面内容语义的过滤。元数据方法需要人们在设计页面时增加许多原来不需要的 Meta 信息，而这一点对设计者要求过高，因此目前此方法并不实用。根据页面语义的过滤，需要对整个文本进行相关度计算，速度较慢，不能符合人们实时性的要求，扩展元数据方法主要是利用链接周围的 Meta 信息来预测所链到的页面主题，尽管可靠性不如根据页面语义方法高，但有较好的实时性。因此，我们的系统采用了综合扩展元数据方法和链接分析方法的 IPagerank 方法。也就是说，我们的方法是进行 URL 与主题的相关性判定。按照高预测值优先采集、低预测值(小于设定阈值)被抛弃的原则进行剪枝处理。这样可以大大减少采集页面的数量，有效地提高主题信息搜索的速度和效率。

3.5.1 扩展元数据的含义：

尽管目前元数据演算(在 HTML 中增加的一类标记，记作<Meta> </Meta>)并不理想，人们却发现利用其它 HTML 标记 anchor 等信息能够有效的指导检索和基于主题的信息采集。为了与元数据相区别，我们把这些标记信息统称为 HTML 扩展元数据，相应的计算叫做扩展元数据演算。整个扩展元数据类型可以分为 3 个大类：1).URL（包括 HREF，OnMouseover，Src 等）；2).Text(包括 Anchor Text，Image Text，Map&Area Text，Frame Text 和 Surrounding Text 等)；3)Title(包括 Title，Name 等)。

3.5.2 扩展元数据启发式算法(All Metadata Heuristics or AMH)

我们发现，如果一个 URL 中包含某个主题词，则这个 URL 所指向的页面很可能是跟这个主题词密切相关的。比如 http://dmoz.org/Sports/Basketball 这个 URL 包含的内容就很可能是关于 Basketball 的。因此定义 URL 启发式算法(URL Heuristics or UH)公式如下：

$$\Theta_{UH}(url)=\begin{cases}1 & \text{如果这个 } url \text{ 中包含主题词}\\ 0 & \text{否则}\end{cases} \qquad \text{公式 1}$$

一般来说，根据这个公式计算的值ΘUH 如果为 1，则这个链接所指向的页面与主题相关的准确性很高，但算的值ΘUH 如果为 0，这个链接所指向的页面与主题无关的准确性并不高。也就是说此算法给许多实际相关的页面并没有赋权值 1。

与 URL 启发式算法类似，还有 Text 启发式算法(Text Heuristics or TeH)和 Title 启发式算法(Title Heuristics or TiH)公式如下：

$$\Theta_{TeH}(url)=\begin{cases}1 & \text{如果这个 } url \text{的} Text \text{中包含主题词}\\ 0 & \text{否则}\end{cases} \qquad \text{公式 2}$$

$$\Theta_{TiH}(url)=\begin{cases}1 & \text{如果这个 } url \text{的} title \text{中中包含主题词}\\ 0 & \text{否则}\end{cases} \qquad \text{公式 3}$$

3.5.3 扩展元数据方法：相关性权重算法(Relevance Weighting or RW)

$$\Theta_{RW}(url)=\begin{cases}\max(\ \Theta(t)_{\,t\in M(url)}\ ) & \text{如果}\max(\Theta(t))\geq c\\ 0 & \text{否则}\end{cases} \qquad \text{公式 4}$$

其中，M(url)指与此 URL 相关的所有扩展元数据集合，$\Theta(t)$ 是指扩展元数据中的一个词与主题的相关度。c 为用户设定的相关性阈值。一般的扩展元数据方法是看扩展元数据中是否包含主题词或者主题词的同义词，这样会漏掉许多相关页面；而 RW 方法则是看扩展元数据中词与主题词之间的相似度，同义词之间的相似度 100%，近义词之间的相似度 50%~100%,远义词之间的相似度 0%~50%，这样大大降低了漏判相关页面的可能性，同时也增加了错判相关页面(不相关的页面判断为相关页面)的可能性，它的相关与否是通过阈值来决定的(大于等于阈值为相关，小于阈值为不相关)。

3.5.4 链接分析方法：PageRank 算法

PageRank 是著名搜索引擎 Google 的一个重要检索算法，它定义如下：给定一个网页 A，假设指向它的网页有 T_1，T_2，…，T_n。令 C(A)为从 A 出发指向其它网页的链接数目，A 的 PageRank PR(A)通过公式 5 计算，其中 d 为衰减因子(通常设成 0.85)。

$$PR(A)=(1-d)+d\left(\frac{PR(T_1)}{C(T_1)}+\frac{PR(T_2)}{C(T_2)}+\cdots+\frac{PR(T_n)}{C(T_n)}\right) \qquad \text{公式 5}$$

3.5.5 IPageRank 算法

尽管 PageRank 方法对发现重要页面有很强的能力，但是它发现的重要页面是针对广泛主题的，而不是基于一个具体的主题。因此，一个被大量无关于主题的页面群指向的页面的 PageRank 值就比一个由少量相关于主题的页面群指向的页面的 PageRank 值高。当然，一个被大量相关于主题的页面群指向的页面的 PageRank 值往往也高于一个由少量相关于主题的页面群指向的页面。因此，我们对 PageRank 方法进行了改进：在链接关系的基础上，加入一定的语义信息权重，以使得所产生的重要页面是针对某一个主题的，这就形成了 IPageRank 算法。IPageRank 算法既利用了 PageRank 发现重要页面的优势，又利用 RW 算法提高链接的

相关性。改进公式如下：

$$IPR(A) = (1-d) + d\left(IPR(T_1) \bullet \frac{\Theta_{RW}(url_{T1})}{\sum_{1}^{k1} \Theta_{RW}(url_i)} + IPR(T_2) \bullet \frac{\Theta_{RW}(url_{T2})}{\sum_{1}^{k2} \Theta_{RW}(url_i)} + \cdots + IPR(T_n) \bullet \frac{\Theta_{RW}(url_{Tn})}{\sum_{1}^{kn} \Theta_{RW}(url_i)} \right) \quad \text{公式 6}$$

其中，A 为给定的一个网页，假设指向它的网页有 T_1，T_2，…，T_n。url_{T1}，url_{T2}，…，url_{Tn}分别是网页 T_1，T_2，…，T_n指向 A 的链接，k_1，k_2，…，k_n分别是网页 T_1，T_2，…，T_n中所含的链接数。IPR(A)为 A 的 IPageRank 值，d 为衰减因子(也设成 0.85)。

IPageRank 的实际意义可以这样来解释。假设 Web 上有一个主题浏览者，IPageRank(即函数 IPR(A))是它访问到页面 A 的概率。它从初始页面集出发，按照页面链接前进，从不执行“back”操作。在每一个页面，浏览者对此页面中的每个链接感兴趣的概率是和此链接与主题的相关性成比例的。当然浏览者也有可能不再对本页面的链接感兴趣，从而随机选择一个新的页面开始新的浏览。这个离开的可能性设为 d。从直观上看，如果有很多页面指向一个页面，那么这个页面的 PageRank 就会比较高，但 IPageRank 值不一定很高，除非这很多的页面中大部分都为与主题相关的页面；如果有 IPageRank 很高的页面指向它，这个页面的 IPageRank 也会很高。

3.6 页面与主题的相关性判定

为了进一步提高采集页面的准确率，需要对已采集的页面进行主题相关性评价，也就是页面过滤。通过对评价结果较低的页面(小于设定的阈值)剔除，来提高所采集主题页面的准确率。我们采取的方法就是基于关键词的向量空间模型算法。

4 系统的实现

4.1 系统基本情况

为了对不同算法进行评测，我们在原有基于站点采集的“天罗”采集系统的基础上改进，构建了“天达”主题 Web 信息采集系统。“天罗”信息采集系统是一个采集性能较高的实用系统，能够高效地采集包括 Web 网页、FTP 文件、Web 聊天、Web BBS、以及 Telnet BBS 等多种信息。

4.2 系统测试结果

我们选择了旅游信息作为主题进行测试，收集了旅游主题网站 20 个，并加入了 60 个无关网站组成测试集，网页总数超过 20000。

我们用相同的初始 URL 集合，分别用宽度优先(RW)ageRank 算法、IPageRank 算法、对数据进行采集。为了有效地得到各个方法的准确效果，我们在实验中暂停了页面与主题相关性判定模块，并及时的计算出采集准确率和资源发现率

比较 / 算法	采集准确率	资源发现率
宽度优先	35%	100%
RW	88%	49%
PageRank	29%	30%
IPageRank	68%	86%

图 3

结果 / 指标	测试结果	评价
最终采集准确率	76%	较高(优点)
最终资源发现率	80%	较高(优点)
内存的占用	30M (估计)	较大(缺点)
采集速度	80 页/分钟	较慢(缺点)

图 4

4.2.3 性能测试

我们的测试平台为一台 CPU 为 Intel PIII 800、内存为 128 兆、操作系统为 Window2000 Professional 的计算机，在采集时候，系统的采集端设置了 10 个线程，采用的 URL 预测算法为 IPageRank。所测试的性能指标包括最终采集页面的准确率、采集页面的资源发现率、内存的占用大小，测试结果如图 4。

5 结束语

随着人们对 Web 服务种类和质量要求的提高，我们展开了基于主题的 Web 信息采集技术的研究，并设计了一个实际系统。在原有技术的基础上，我们又设计出许多独具特色的新算法，比如 Spider 采集、标题提取、URL 主题预测以及页面与主题相关性的判定，特别地，我们对著名的 Google 算法进行了改进，以使得它即适合基于主题的采集，又保持了原来的优势。实验表明基于主题的采集优势是明显的。随着 Web 服务朝个性化方向的迈进、Agent 技术的发展、迁移式思想的出现，单纯的为了检索的 Web 信息采集技术必将向着基于主题以及个性化主动信息采集服务方向全方位拓展。

参 考 文 献

[1] [Aggarwal et al. 2001] C. Aggarwal, F. Al-Garawi and P. Yu. "Intelligent Crawling on the World Wide Web with Arbitrary Predicates". In Proceedings of the 10th International WWW Conference, May 2001.

[2] [Brin & Page 1998] S. Brin and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In Proceedings of the Seventh International World Wide Web Conference, Australia, April 1998.

[3] [M.Diligenti et al. 2000] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori Focused Crawling Using Context Graphs. VLDB Conference. 2000

[4] [Cho et al. 1998] "Efficient Crawling Through URL Ordering", J. Cho, H. Garcia-Molina, L. Page. In *Proceedings of the 7th International WWW Conference*, Australia, 1998

[5] [Diligenti et al. 2000] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori Focused Crawling Using Context Graphs. VLDB Conference. 2000

[6] [Hersovici et al. 1998] "The Shark-Search Algorithm - An Application: Tailored Web Site Mapping", M. Hersovici, M. Jacovi, Y. Maarek, D. Pelleg, M. Shtalhaim and S. Ur. In *Proceedings of the Seventh International World Wide Web Conference*, Australia, April 1998.

[7] [Kleinberg 1998] J.Kleinberg.Authoritative Sources in a Hyperlinked Environment.SODA.1998

[8] [Marchiori 1998] Massimo Marchiori, "the limits of web metadata, and beyond", proceeding of the 7th World Wide Web Conference, 1998.

[9] [Selberg&Etzioni 1995] Erik Selberg and Oren Etzioni. "Multi-Service Search and Comparison Using the MetaCrawler". In Proc. 4th World Wide Web Conference, Boston, MA USA, December 1995

[10] [SMB 1999] Soumen Chakrabarti,Martin van den Berg,Byron Dom,"Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery". Proceedings of the 8th World-Wide Web Conference,1999,Canada.

[11] [李盛韬 2002] 李盛韬. Web 信息采集研究进展 计算机科学, 2002.

[12] [余智华 1999] 余智华. "WWW 站点的分析与分类" 硕士论文[D]. 北京：中科院计算所 1999