第 33 卷 第 6 期 计 算 机 工 程 2007 年 3 月
Vol.33 No.6 Computer Engineering March 2007
·软件技术与数据库· 文章编号：1000—3428(2007)06—0080—03 文献标识码：A 中图分类号：TP311

WWW 论坛中的动态网页采集

李 魁[1,2]，程学旗[1]，郭 岩[1]，张 凯[1]
（1. 中国科学院计算技术研究所，北京 100080；2. 中国科学院研究生院，北京 100039）

摘 要：网络论坛已经成为互联网信息发布的主要形式，对论坛信息的检索和挖掘都涉及到论坛信息的获取，然而传统的针对静态网页的广度优先采集工具，不能有效地获取论坛信息。该文利用论坛的结构特点，提出了一种“版面-主题关联判断”(BTCJ)算法，采用一种基于版面扩展的采集策略。实验证明，该方法在论坛采集准确率和覆盖率方面显著优于广度优先策略；具有良好的泛化能力，应用在实践中已覆盖各种类型的论坛 12 000 余个。

关键词：互联网论坛；信息采集；动态网页

Crawling Dynamic Web Pages in WWW Forums

LI Kui [1,2], CHENG Xueqi [1], GUO Yan [1], ZHANG Kai [1]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;
2. Graduate School, Chinese Academy of Sciences, Beijing 100039)

【Abstract】 Web Forums have been one of dominating ways for information release and exchange in Internet. Crawling is the groundwork of searching and mining information from Web Forums. However, traditional crawling component usually using “Broad-first” strategy can not fetch information from Web Forums effectively. Exploring inner structure-features of forums, this paper presents a crawling strategy, which is based on “board-topic correlation judgments” algorithm. Compared with “board-first” strategy, this solution performs remarkably better both in precisions and recall. In practice, the algorithm is performed over 12 000 different Web forums and achieves a good result.

【Key words】 WWW forums; Information crawling; Dynamic Web page

WWW 论坛对比一般的网站具有交互性、参与性、内容新颖、涉猎面广的特点，已经成为现时代一种非常热门的信息获取渠道。论坛中蕴涵着大量富有使用价值和商业价值的内容，挖掘网上论坛不仅可以便利广大网民搜索网络资源，也可以提供给第 3 方公司有意义的信息资料。同时由于论坛反映的是和用户密切相关的内容，论坛成为互联网上一个非常重要而独特的信息宝库，针对于 WWW 论坛的信息采集具有越来越重要的意义。

不同于普通网站以静态网页为主，WWW论坛大多借助于数据库和动态网页技术生成，使得传统的采集方式遇到了前所未有的困扰：采集陷阱和采集质量低下[1~3]。作为应对，当前流行的搜索引擎往往采用消极的规避策略，尽量避免过多采集论坛中的动态页面[4,5]，这使得WWW论坛中的资源不能得到有效的采集利用。

1 WWW 论坛的特点与采集的难点

对于 WWW 论坛，论坛中的链接具有如下独特的性质：

(1)链接的种类多，除了访问资源的超链接，还存在大量功能性的链接和噪声链接，所谓功能性的链接，即完成某种特定操作的链接，如“发表”,“评论”等功能。

(2)链接的层次深，大量的内容需要深入论坛才能访问到。

(3)链接冗余现象明显，所谓链接冗余即指同一内容存在多个不同的链接与之相对。

这些特点使得广度优先的采集策略在论坛采集中受到了严峻的挑战，一方面采集的负担明显加重，容易陷于采集陷阱，消耗大量的资源；另一方面采集的效率非常低下，大量无意义、重复的链接被采集。

由于以上原因的存在，因此在对动态网页的采集中，采集陷阱是一个非常棘手的问题。所谓采集陷阱，是指采集器陷入在网站的链接的无穷尽的扩展中，对采集器而言，此网站被认为存在无穷的链接需要被采集。采集陷阱的存在，会导致采集器有限的资源的白白浪费，甚至使采集程序崩溃。为了避免陷入采集陷阱，当今主流搜索引擎对动态网页的采集都采用相应的规避策略：如限制采集单个网站的层次和限制采集的数量。另外，大量噪声链接和冗余链接的存在，使得采集的精度大大降低，大量无实际意义的网页被采集。

2 基于版面扩展的论坛采集策略

分析互联网论坛的结构，可以发现互联网论坛中存在“版面-主题索引页-主题”的 3 层扁平逻辑结构。如图 1 所示，论坛被人为地根据不同的话题类别组织成若干个讨论区，称之为版面。用户在讨论区中对相关话题发表讨论，用户发表的一个主题的内容称之为主题，一个主题包括其后的跟帖。版面是同类型主题的集合。主题是采集器在论坛中唯一关心的信息资源，从论坛首页出发，找到各个版面，再获得版面中的所有主题是一种很自然的想法。

进一步发现，在各个版面中，主题是以列表的形式分页呈现，若干个主题的列表形成一个分页，同一版面的各分页之间通过“上一页”,“下一页”之类的链接相互链接。因分页中含有大量主题的链接，把这样的分页称为主题索引页面。

基金项目：国家“973”计划基金资助项目“大规模文本内容计算”(2004CB318109)
作者简介：李 魁(1982－)，男，硕士生，主研方向：信息检索，自然语言处理；程学旗，研究员；郭 岩、张 凯，助理研究员
收稿日期：2006-03-25 **E-mail：**ibucan@126.com

版面的所有主题页面的集合组成了该版面主题集合的一个分划。主题索引页面是从版面得到主题的中介。

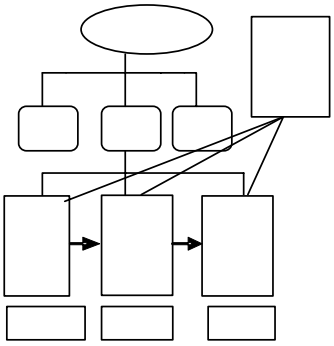

图 1 论坛逻辑结构

在采集的过程中利用论坛的逻辑结构，通过主题索引页面定位主题，实现对论坛的精确采集。首先找到论坛的各个版面；再从版面开始得到该版面的所有主题索引页面；最后从主题索引页面中提取各个主题的链接。称之为基于版面扩展的采集策略。

3 算法描述

3.1 链接的分类

论坛采集中面对的主要是动态网页的链接。所谓动态网页是指网页在服务器端并不存在，而是在客户请求时由服务器端即时生成的网页。动态网页对应的 URL 称为动态 URL，动态 URL 中含有服务器端脚本，需要传递的参数名和参数值，Web 服务器在得到对动态 URL 的请求后，解析相应服务器端脚本，并结合参数生成网页内容发送给客户端。

如下为动态 URL 的示例：

http://ServerHost/Dir/ScriptName?Para1=Value1[Para2=Value2....]

其中，ServerHost 为脚本所在站点域名，Dir 为脚本所在 Web 服务器目录路径，ScriptName：脚本名称，Para(i) 为传递给脚本的第 i 个参数的名称，Value(i) 为传递给脚本的第 i 个参数的值。

链接分类的思想在于：在同一站点中，性质相似的网页具有相似的 URL。WWW 互联网论坛网页中的链接按照性质可以分为若干类，如主题链接、版面链接、显示用户信息的链接、其它功能性链接等。这些不同类型的链接在 URL 上表现都有自己的特征。这种特征在单个 URL 上不一定能够清晰地体现出来，但当同种类型的链接的 URL 作为一个整体的时候，这种 URL 内在蕴涵的特征就能呈现，明显地区别于其它类型链接的 URL。

我们这样定义动态 URL 的相似性：同一站点中，具有相同的 ScriptName，相同的参数名和相同的参数个数的两个 URL 是相似的。

以此作为动态链接分类的依据，使不同类型的链接得到区分，在试验中，这种链接的分类方法在互联网论坛链接分类的应用上具有非常良好的分类精度和区分性，而且分类结果具有良好的可解释性。

3.2 版面链接的判断

从论坛的首页中抽取得到站内链接，按照上面的链接分类算法对链接进行分类，得到若干个链接类，排除过小的类，得到候选版面链接类。此时，确定版面链接类有需要描绘版面链接类的特征。

我们提出了一种版面主题关联判断算法来确定版面链接类。版面链接指向的页面为版面页面，通透过判断链接指向的页面是否为版面页面来判定链接是否是版面链接。版面页面实际上是特殊的主题索引页面，它是该版面中的第一个主题索引页面，其中内容主要是帖子列表显示。从以下两方面去归纳版面页面的特征：

(1)链接描述文字特征

版面页面中的主体是主题列表显示，其中包含一定数目主题的链接集合，这些链接对应的链接描述文字则是关于主题内容的概括。我们发现主题的链接描述文字通常具有完整的语义，这从长度上反应出来是具有较大的长度，区别于版面页面中其它链接。

(2)数量特征

版面页面的主题列表的分页显示，这决定了主题链接在版面页面具有相当数量。

版面-主题关联判断算法(BTCJ)如下：

(1)提取待判断页面中的链接；

(2)对所有站内链接使用前述分类算法分类；

(3)若链接类满足这样的条件：类的大小超过一定值，且此类的锚文本平均长度大于阈值，则标注此类为主题链接类；

(4)当且仅当存在单一的主题链接类时，此页面被认为是版面页面，否则不是。

3.3 主题索引链接的自动翻页扩展

找到了站点的版面链接后，还要获得各个版面中所有的主题索引页面的链接。如上文所述，版面页面为该版面的第 1 个主题索引页面，称之为种子索引页面。

同一版面内的各个主题索引页面之间通过链接相互连接。从版面种子索引页面出发，通过链接蔓延再加上一些启发式信息可能得到该版面的所有主题索引页面。但这不是一种经济的方法，若版面的主题索引页面较多即意味着要深入很多层链接才能得到所有的主题索引页面，链接蔓延的代价随深度呈指数增长。另一方面，主题索引页面具有的逻辑先后顺序在链接蔓延中丢失，这对更新是不利的。

我们已经得到版面链接，这是一种特殊的主题索引页面链接，它与该版面的其它主题索引页面链接在 URL 的表现形式上存在相关性。需要利用这种相关性找到该版面的其它主题索引链接的样例。我们注意到，同一版面所有的主题索引页面链接(除种子索引页面)的 URL 上具有相似性，变化的只是个别参数的值，这个参数值的不同对应着该版面不同的主题索引页面链接，此参数称之为翻页参数。相邻两个主题索引页面的翻页参数值之间的差值，称之为翻页参数间距。

http://bbs.myadobe.com.cn/forumdisplay.php?f=85&page=5&sort=lastpost&order=&pp=20&daysprune=-1

http://bbs.myadobe.com.cn/forumdisplay.php?f=85&page=6&sort=lastpost&order=&pp=20&daysprune=-1

上例所示翻页参数为 page，翻页参数间距为 1。确定了翻页参数和翻页间距，就可以方便地得到该版面的所有主题索引页面。最后通过这些主题索引页面，就能得到论坛中有效信息(主题)的链接。

4 实验

4.1 WWW 论坛中的“采集陷阱”现象

以对一组 WWW 论坛站点以广度优先策略进行采集，采集深度为 5。记录每个站点实际采集网页的数量，与该站点中实际所有的主题数进行比较。

从表 1 可看出广度优先策略在论坛采集中遇到的采集陷阱问题：采集器获取的页面数远远大于论坛中实际的主题数。

表 1 广度优先采集数与帖子数对比

论坛站点名	BF—5 采集网页数	实际主题数
WWW.cntong.com/phpbb/	94 193	3 452
csapa.org/phpBB/	155 560	7 201
forum.lnnu.edu.cn	827 318	78 388
bbs.centrmus.com	616 286	16 435

4.2 与广度优先算法的比较

在 WWW 论坛中，对采集器而言有效的网页是用户在论坛中发表的主题。定义论坛

采集的准确率和覆盖率：

准确率=采集得到主题数/采集的网页总数

覆盖率=采集得到主题数/论坛主题数

选取了各种类型的 WWW 的论坛，对广度优先算法和我们的算法(BPCJ)的采集准确率和覆盖率进行了比较，结果如表 2 所示。

表 2 论坛采集算法与广度优先采集算法效果比较

站点名称	论坛类型	论坛主题数	采集的网页总数	BTCJ 算法			广度优先采集算法		
				采集得到主题	采集准确率%	采集覆盖率%	采集得到主题	采集准确率%	采集覆盖率%
csapa.org	phpbb	7 201	6 601	6 554	99.3	91.0	2 219	33.6	30.8
WWW.cntong.com	phpbb	3 452	2 531	2 472	97.7	71.6	1 352	53.4	39.2
WWW.linuxbyte.net	IBP	3 020	2 148	2 097	97.6	69.4	633	29.5	21.0
WWW.soking.com	BMForum	2 605	1 949	1 849	94.9	71.0	357	18.3	11.8
luntan.popo.163.com	自设计	142 315	50 378	49 122	97.5	34.5	3 597	7.0	2.5
envi.ruc.edu.cn/bbs/	Newvbb	2 103	1 899	1 713	90.2	63.4	326	17.2	15.5

BPCJ 算法的采集准确率达到了 90%以上,明显优于广度优先算法，采集覆盖率也比广度优先算法有了显著提高，个别站点采集覆盖率较低的原因是该站点中主题数目众多，而 BPCJ 扩展的版面翻页数有限(在实验中扩展翻页数为 32)。

4.3 结论

从表 2 中可以看出，我们的算法在对论坛中的动态网页采集上具有明显的优势。通过对版面和主题链接的识别，以及扁平化的采集策略解决了互联网论坛中采集陷阱这一阻碍采集的根本问题，消除了噪声链接造成的采集质量低下的困扰。而值得一提的是，我们的算法并不针对某一论坛或某一类型的论坛，它不需要训练学习，也不需要对某一站点指定特定规则，而是总结了互联网论坛在逻辑上和应用技术上的内在规律，具有极强的泛化能力。在实际中已经覆盖了各种类型的论坛站点 12 000 余个，这进一步证明了我们先前的假设：在同一站点中，功能相似的链接在 URL 形式上具有很强的相似性。

参考文献

1 Cho J, Garcia-Molina H, Page L. Efficient Crawling Through URL Ordering[C]//Proceedings of the 7th International World Wide Web Conference. 1998: 161-172.

2 Najork M, Wiener J L. Breadth-first Crawling Yields High-quality Pages[C]//Proceedings of the 10th International World Wide Web Conference. 2001: 114-118.

3 Li Jun, Furuse K, Yamaguchi K. Focused Crawl -ing by Exploiting Anchor Text Using Decision Tree[C]//Proceedings of the 14th International World Wide Web Conference. 2005: 1190-1191.

4 Castillo C. Effective Web Crawling[D]. University of Chile, 2004.

5 Brin S, Page L. The Anatomy of a Large-scale Hypertextual Web Search Engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.

~~~~~~

(上接第 73 页)

遗忘协同过滤算法的 MAE 值都小于传统的协同过滤算法，但是对于不同的 k，线性逐步遗忘协同过滤算法的优势不一样，显然它受邻居用户大小 k 的影响。

3.4 结论

由前面的实验结果可得到如下结论：总体上线性逐步遗忘协同过滤算法在准确性方面优于传统的协同过滤算法。由于用户兴趣大多都是逐渐改变的，因此在推荐系统中使用线性逐步遗忘策略将有效提高推荐算法的准确性。

4 结束语

针对协同过滤系统中的用户兴趣变化问题，本文提出了线性逐步遗忘协同过滤算法。实验结果表明，当用户兴趣发生变化时，线性逐步遗忘协同过滤算法在准确性方面优于传统的协同过滤算法。

参考文献

1 Karypis G. Evaluation of Item-based Top-N Recommendation Algorithms[R]. Minneapolis: Dept. of Computer Science, University of Minnesota, Technical Report: #00-046, 2000.

2 Cai Deng, Lu Zen-xiang, Li Yanda. Collaborative Filtering[J]. Computer Science, 2002, 29(6): 1-4.

3 Sarwar B, Karypis G, Konstan J. Item-based Collaborative Filtering Recommendation Algorithms[C]//Proc. of the 10th International World Wide Web Conference. 2001: 285-295.

4 Deshpande M, Karypis G. Item-based Top-N Recommendation Algorithms[J]. ACM Transactions on Information Systems, 2004, 22(1): 143-177.

5 Goldberg D, Nichols D, Oki B M. Using Collaborative Filtering to Weave an Information Tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70.

6 Koychev I, Schwab I. Adaptation to Drifting User's Interests[C]//Proc. of ECML'00, Barcelona, Spain. 2000.

7 Kukar M. Drifting Concepts as Hidden Factors in Clinical Studies[C]//Proc. of the 9th Conf. on Artificial Intelligence in Medicine in Europe, Protaras, Cyprus. 2003.

8 Zeng Chun, Xing Chunxiao, Zhou Lizhu. A Survey of Personalization Technology[J]. Journal of Software, 2002, 13(10): 1952-1961.

9 Yu Li, Liu lu, Li Xuefeng. Research on Personalized Recommendation Algorithm for USer's Multiple Interests[J]. Computer Integrated Manufacturing Systems, 2004, 10(12): 1610-1615.

10 Yuan Genqing. Medical Psychology[M]. Nanjing: Southeast University Press, 1995.
~~~~~~