doi:10.3969/j.issn.1673-4785.2010.02.003

Web 页面信息主动检索模型

袁鼎荣[1,2],钟 宁[1]
(1.北京工业大学 国际 WIC 研究院,北京 100022; 2.广西师范大学 计算机科学与信息工程学院,广西 桂林 541004)

摘 要:单个页面信息量远远大于特定用户对页面中的信息需求.为快速准确从当前页面中获取特定用户所需求的兴趣信息,提出了页面信息主动检索模型.该检索模型中,根据页面 Block 特点将当前 Web 页面转化成信息树,根据用户过去的浏览行为构造用户特征树,挖掘用户特征树产生用户需求信息集,然后从当前页面中检索需求的信息,获取用户兴趣信息集.详述了主动检索的基本原理,给出了相应的算法描述,并通过实验证明了该模型具有可行性.

关键词:页面 Block;页面信息树;用户特征树;主动检索
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-4785(2010)02-0112-05

Initiative retrieval of web information

YUAN Ding-rong[1,2], ZHONG Ning[1]
(1. The International WIC Institute, Beijing University of Technology, Beijing 100022, China; 2. College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China)

Abstract: The information capacity of a single web page is far more than the needs of any individual user. The goal of the authors was to construct a discriminative means for retrieving individualized web page information. Such a model could allow fast and precise retrieval of information from current pages, tuned as required by particular individuals. The approach of the researchers was to begin by transforming a web page into an information tree in light of the block characteristics of the web. Next, a user characteristics tree was constructed from user behavior seen in previously browsed pages. This was mined to get information needed about the user. Based on this, elements interesting to the user were retrieved from current pages. The basic principles of discriminative retrieval were introduced, several retrieval algorithms described, and the model's feasibility experimentally verified.

Keywords: Web Block; page information tree; user especial tree; initiative retrieval

信息检索是指依据一定的方法,从已经组织好的大量的信息集合中,查找并获取特定相关信息的过程.传统意义上的信息检索由用户、搜索引擎和数据集组成.以引擎为中心,由搜索引擎根据用户的信息需求展开检索,然后将索引结果反馈给用户.搜索的数据源包括关系数据库、文本数据集和 Web 上的多媒体数据.这种检索技术在关系数据库和文本数据集的查询检索中取得了良好的应用.不管成熟的数据库查询还是富有挑战性的 Web 信息检索都是以搜索引擎为中心,当用户提出检索要求时,搜索引擎才依据用户要求检索后台数据,最后将检索结果返回给用户.随着 Web 技术的发展,网络应用的普及,网络信息资源急剧增长,页面信息的无限性与用户接受信息的有限性的矛盾日益突出.尤其手机用户接入 Internet 后加剧了这种矛盾.因此如何从当前页面中快捷准确地发现特定用户潜在兴趣的有限信息集成为新的应用挑战,为解决这种矛盾提出一种称为主动检索的检索模型,即搜索引擎根据用户特征,自动产生检索条件,主动检索当前页面,提取用户潜在兴趣的页面信息反馈给用户.

收稿日期:2009-12-04.
基金项目:国家自然科学基金重大研究计划资助项目(90718020);澳大利亚 ARC 资助项目(Australian Research Council Discovery Grant, DP0667060).
通信作者:袁鼎荣. E-mail:dryuan@mailbox.gxnu.edu.cn.

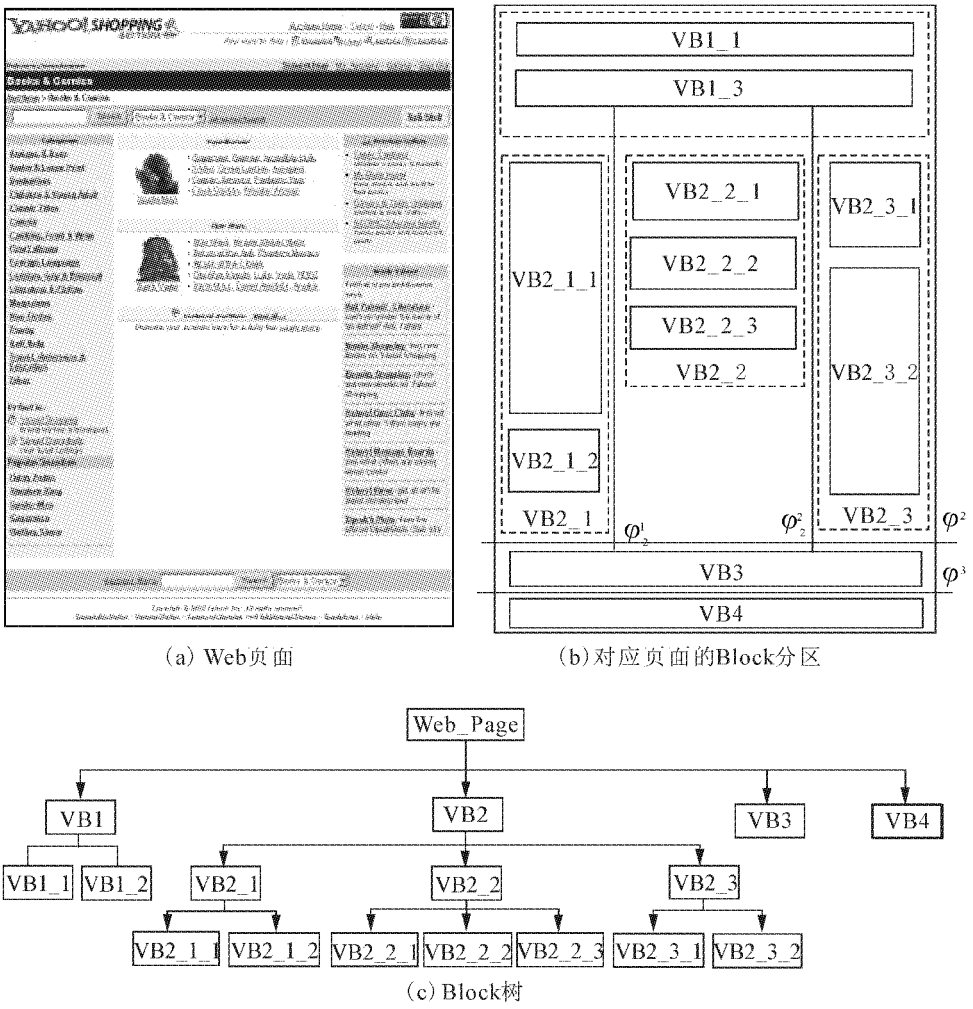
1　相关研究背景

1.1　DOM 树

DOM 即 document object model,它是 W3C 推荐的用于访问诸如 XML 和 XHTML 文档的标准. DOM 把 XML 和 XHTML 文档视为一棵树,文档中的各组成部分定义为一个节点,节点主要包括元素节点、属性节点和文本节点等,这棵节点树展示了节点的集合,以及它们之间的联系,从根节点开始,然后在树的最低层级向文本节点长出枝条,节点之间的父子关系表示标签在页面之中的嵌套关系,兄弟关系表示了标签在页面中的并列关系,这样在使用 DOM 树进行 Web 页面解析时,将所有的页面中的元素都解析成树中的节点,最后通过解码器将解析的结果以语法树的形式输出,内存中生成 DOM 标记树,实现对整个文档的全面的、动态的有层次的访问,每一个网页对应一个 DOM 标记树.

1.2　页面 Block 树

页面 Block 是页面中在内容或显示上独立的、闭合的矩形区域. Web 页面可以分割为若干个互不相交的 Block,把这个过程称为页面 Block 分区. 一个 Block 可以由多个相互不重叠的子 Block 组成,如图 1(a)为 Web 页面,图 1(b)为对应页面的 Block 分区,图 1(c)为依据图 1(b)的分区结构和层次所对应的树形结构,称为 Block 树.



(a) Web页面　　(b)对应页面的Block分区

(c) Block树

图 1　页面分区和页面 Block 树的示例

Fig. 1　The example of page subarea and Web Block tree

分区级别表示该分区产生的 Block 树的最大深度,用 Level 表示. 选择合适的分区级别有助于得到最理想的分区结果. Block 可以定义成一个由主题、内容组成的二元组:Block = < topic, content >,其中,topic 为 Block 主题,content 为 Block 内容. 用 Layout 表示网页的布局.

1.3　页面结构分析

DOM 提供了树形结构的页面模型,可以基于 DOM 来建立 Block 树,由于 HTML 语法的灵活性,很多 Web 页面并不完全遵守 W3C 的 HTML 规范,

依据 DOM 的分区只能代表布局结构独立的区域而不能完全代表语义上独立的区域. 比如,同一个父节点下面的 2 个子节点并不一定代表相同的主题,因此,进行页面结构分析时不能完全依赖 DOM,还需要考虑其他一些因素. 比如借助页面中大量的 HTML 标签来获取布局和位置信息,如 < P > 、< TABLE > 、< TR > 、< UL > 、< H1 > ~ < H6 > 等. 通常,显示上独立的区域一般表示相同的主题, Web 中提供了大量可见的元素来划分页面,例如字体、颜色、图像、空白,这些都是页面 Block 分区算法需要考虑的元素.

分区过程中,从页面开始,逐级递归分区,抽取出第 n 级 Block(n 初始化为 1),组成深度为 n 的 Block 树,同时保存 Layout. 然后,判断 n 是否满足给定的分区级别,如果小于该级别,则依次把 Block 树中的每个叶子节点 Block 的内容(content)作为新的 DOM,重新抽取 Block 并组装 $n+1$ 级 Block 树,如果某个 Block 已经无法再分割,则忽略这个 Block. 如此反复,直到 Block 树达到给定分区级别为止. 对于页面分块的详细规则见文献[1-4],这里不再重复.

图 1(a)例子页面通过 3 次分区算法迭代,产生了图 1(c)的 Block 树.

2 主动检索模型

Web 页面充斥着形形色色的信息,从信息的表现形式来看有文本、动画、声音和视频等;从页面信息的布局来看,有纯内容页面、目录页面,也有介于两者之间的综合性页面,如新浪、腾讯等[5-10]. 为简化起见,以综合性页面为研究对象.

在检索模型中,首先对页面数据进行预处理,根据页面的 Block 特点将平面的 Web 数据转换为对应的树,称这种树为 Web 查询树,简称 Web 树.

2.1 Web 树的建立

首先将 HTML 代码转换为 XML 并生成 Dom 树,借助页面分区标记将页面分成逻辑上不同的 Blcok,再结合 DOM 树将页面信息用树形结构表示,然后对所获得的树进行清理剪枝,获得 Web 树. 算法描述如下.

算法 1: Web 树构造算法.

Input: html; //输入页面代码文件

Output: Web Tree. //输出信息检索树

1) xml file = htmlparser(html); //将 html 文件格式解释为 xml 代码.

2) Domtree = xlmparser(xml file); //将 xml 文件代码解释为一棵文档对象树.

3) Blocktree = getblcoktree(html file);//从逻辑上将页面分成具有一定等级和关联的 Block, 并将其转换成树形结构.

4) Webtree = merge(domtree, Block tree);//根据 DOM 树和 Block 树构造初步的 Web 树

5) Webtree = pruning(Webtree);

6) output Webtree.

如图 2 为一棵 Webtree.

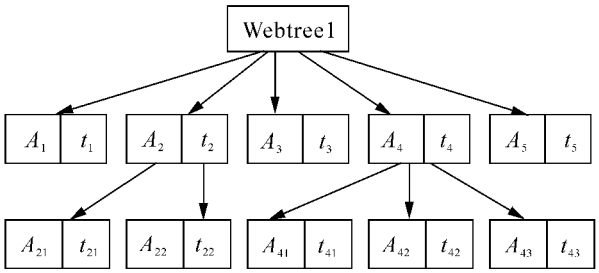


图 2 页面树

Fig. 2 Webree

该树有 5 个儿子结点,代表 5 个不同的页面板块, 其中 A_2 板块包括 2 个子版块, A_4 板块包括 3 个子版块.

2.2 用户特征树

定义结点为: node = <topic, content, frequent>, 其中, topic、content 对应 Webtree 中的结点, frequent 为频度,即访问 Webtree 中相应结点的次数. 根结点对应特定的用户. 如果某用户访问页面中的某一条目,则 Webtree 中形成一条路径,称为主题路径. 由特定用户的浏览历史中所记录的主题路径构造用户特征树的算法如下.

算法 2:用户特征树生成算法.

Input: visiting path set; //用户访问的路径集

Output: clienttree;//用户特征树

1) Node = open a new node;

2) Path = getpath(visiting item set);//用户访问的路径集取一路径

3) If clienttree is NULL Then open a new tree and label the tree;

4) clienttree = insert(clienttree, path);//根据用户新的访问记录更新用户特征树, 插入时重叠的结点的 frequent 域加 1;

5) output clienttree.

用户特征树是对用户浏览历史进行分析加工所产生的特定用户的行为特征树,它记录了用户对网络页

面中不同主题、不同板块、不同信息条目的兴趣度.

2.3　潜在页面兴趣信息的近似检索

潜在兴趣信息的检索思想可描述为:当用户启动 Web 浏览器时,首先根据特征树产生潜在兴趣主题向量集 $\boldsymbol{Q}=\{\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_i,\cdots,\boldsymbol{q}_n\}$. $\boldsymbol{q}_i$ 为潜在兴趣主题向量,即用户特征树中的兴趣路径,作为检索条件. 然后依据一定规则,计算 $\boldsymbol{q}_i$ 与当前页面树中各路径的匹配程度,依据匹配度高低选取符合兴趣度要求的页面信息作为检索结果返回给用户.

其算法描述如下:

算法 3:潜在兴趣信息的检索算法.

input: clienttree, webtree, N;

output: InterstingSet;

1) Q = MiningintrestingPath(clienttree, N); //从用户特征树中提取用户兴趣信息集.

2) T = ExtractItemPath(webtree);//从 webtree 中抽取页面信息所在路径.

3) InterstingSet = { };//初始兴趣信息集为空.

4) retrieval (Q, T); //从 webtree 中检索用户潜在 兴趣信息集.

For i = 1 to $|Q|$

For j = 1 to $|T|$

$$\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{T}_j)=\frac{\boldsymbol{q}_i\times\boldsymbol{T}_j}{|\boldsymbol{q}_i||\boldsymbol{T}_j|},$$

If $\mathrm{sim}(\boldsymbol{q}_i,\boldsymbol{T}_j)\geqslant\theta$//$\theta$ 为给定的阈值

Then InterestingSet $\leftarrow\boldsymbol{T}_j$;//将兴趣路径并入兴趣集.

5) output InterestingSet.

3　实验说明

设有用户特征树如图 3,则该用户的主题信息向量集如表 1,如果设置支持度为 7 则有用户的兴趣主题向量集如表 2.

表 1　用户主题向量表

Table 1　Topic vectors table of user

主题向量	频度
B_1	16
B_1B_{11}	3
B_1B_{12}	5
B_2	9
B_{21}	5
B_3	27
B_{31}	7
B_{32}	12

表 2　用户兴趣主题向量表(N = 5)

Table 2　Interesting topic vectors table of user(N = 5)

兴趣主题	频度
B_3	27
B_1	16
B_{32}	12
B_2	9
B_{31}	7

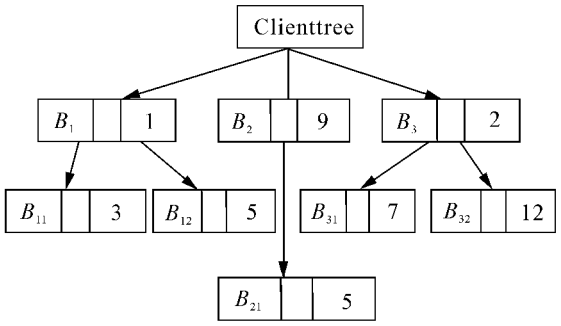


图 3　用户特征树,该树的 3 个儿子结点表示用户有 3 个不同的兴趣主题,分别为 B_1、B_2、B_3

Fig. 3　UC_Tree. There are three son nodes in the tree, which is three different interesting topics B_1, B_2, B_3

设用户访问某主页,该主页所对应的索引树为图 2,则该页面的的主题信息向量集为:$\{A_1,A_2A_{21},A_2A_{22},A_3,A_4A_{41},A_4A_{42},A_4A_{43},A_5\}$ 然后依据式(1)计算用户兴趣信息向量与页面主题信息向量的近似度, 如表 3.

表 3　用户兴趣主题与页面信息向量相似度表

Table 3　Similarity table between interesting topic and webpage information vector

	B_3	B_1	B_{32}	B_2	B_{31}
A_1	0.10	0.20	0.31	0.16	0.30
A_2A_{21}	0.80	0.30	0.25	0.22	0.32
A_2A_{22}	0.10	0.41	0.90	0.10	0.10
A_3	0.20	0.42	0.36	0.30	0.20
A_4A_{41}	0.30	0.32	0.27	0.23	0.00
A_4A_{42}	0.21	0.88	0.32	0.31	0.10
A_4A_{43}	0.22	0.23	0.22	0.46	0.30
A_5	0.18	0.35	0.49	0.26	0.20

如果给定相似度阈值为 0.5, 则有:A_2A_{21}、A_2A_{22}、A_4A_{42} 是用户潜在兴趣信息项.

4　结束语

本文提出页面信息主动检索技术模型,首先给出主动检索技术的定义,然后依据页面信息的特点将 Web 页面转化为树形结构,构造页面信息树. 通

过用户对页面的访问历史事件构造用户特征树,基于用户特征树挖掘用户的浏览行为特征,主动产生针对特定用户的兴趣信息集,以该信息集作为检索条件,检索信息树达到检索页面信息的目的.该技术的发展应用将减少用户从大量繁杂的页面信息中挑选自己感兴趣的数条或数十条信息的劳累,尤其解决了手机用户中,页面信息容量大和显示屏幕极其有限的矛盾.该技术的成熟发展需要良好的文本分类及其文本主题提取技术、页面 Block 主题信息的提取等技术.用户特征树为 Web 用户行为特征挖掘提供良好的技术支持.

参考文献:

[1]CAI D, YU S, WEN J R, MA W Y. VIPS: a version-based page segmentation algorithm MSR-TR-2003-79 [R]. [s.l.],2003.

[2]SONG Ruihua, LIU Haifeng, WEN Jirong, et al. Learning block improtance models for web pages[C]//The 13th International Conference on World Wide Web. New York, USA, 2004:203-211.

[3]CAI D, YU S, WEN J R, et al. Block-based Web search [C]//27th Annual International ACM SIGIR Conference on Information Retrival. Sheffield, UK,2004: 456-463 .

[4]CAI D, YU S, WEN J R, et al. Block-based link analysis [C]//27th Annual International ACM SIGIR Conference on Information Retrival. Sheffield, UK, 2004:440-447 .

[5]宋 杰,王大玲,鲍玉斌,等. 基于页面 Block 的 Web 档案采集和存储[J]. 软件学报,2008,19(2):275-290.
SONG Jie, WANG Daling, BAO Yubin, et al. Collecting and storing web archive based on page block[J]. Journal of Software, 2008, 19(2):275-290.

[6] CHRISTOPHER D. Introduction to information retrieval [M]. England: Cambridge University Press,2009:25-28.

[7]CHEN K J, MA Weiyun. Unknown word extraction for Chinese documents [C]//19th International Conference on Computational Linguistics. Taipei, China,2002:169-175.

[8] HOBBS J R. Information extraction from biomedical text [J]. Journal of Biomedical Informatics,2002,35(4):260-264.

[9]KONGACHANDRA R, KIMPANT C, SUWANAPONG T, et al. Newly-born keyword extraction under limited knowledge resources based on sentence similarity verification[J]. IEEE International Symposium on Communications and Information Technology,2004,21(3):1183-1187.

[10]GAO Junbo, LUAN Cuiju, WANG Xiaofeng. New keyword extraction research[J]. Computer Engineering and Design,2008,29(3):765-767.

作者简介:

袁鼎荣,男,1967 年生,副教授,主要研究方向为文本信息处理、网络智能、机器学习、数据挖掘等.主持或主要参与国家或省部级项目4项,发表学术论文20余篇.

钟 宁,男,1956 年生,教授,博导,主要研究方向为网络智能、知识发现与数据挖掘、粗糙集(Rough Set)与软计算、智能 Agent 技术与应用、脑信息学等,发表学要论文多篇.