第 34 卷 第 3 期 计 算 机 工 程 2008 年 2 月
Vol.34 No.3 Computer Engineering February 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)03—0205—02 文献标识码：A 中图分类号：TP301

基于 WAP 的移动搜索模型

陈 明，孙丽丽
(中国石油大学计算机科学与技术系，北京 102249)

摘 要：针对在 Internet 环境下进行移动搜索的特点，提出一种基于无线应用协议(WAP)的移动搜索模型。该模型利用现有的搜索引擎实现对 Web 信息的检索，并充分利用 Web 页各分割块的特征抽取出与查询请求最为相关的主题区域代替原始网页，以 WAP 协议的标记语言格式输出给用户，方便移动用户快速准确地获取 Web 信息。

关键词：移动搜索；无线应用协议；网页分割

Mobile Search Model Based on Wireless Application Protocol

CHEN Ming, SUN Li-li
(Department of Computer Science and Technology, China University of Petroleum, Beijing 102249)

【Abstract】 Aiming at the characteristics of mobile search under the environment of Internet, this paper puts forward a mobile search model based on WAP. It makes use of existing search engines to get Web information, and makes full use of features of blocks of Web page to extract the theme block which is closely related to query request as the substitute for the primal page content, which is helpful to mobile users' getting Web information on the move.

【Key words】 mobile search; Wireless Application Protocol (WAP); page segment

移动搜索引擎作为搜索引擎发展的新方向，能够满足互联网用户随时随地获取信息的需求，但也给搜索引擎系统的分析、设计和实现带来了新的挑战：(1)移动搜索引擎需要通过无线网络访问 Internet。无线网络的带宽相对有线网络窄很多，目前 Internet 上的网页大多是服务于有线网络环境的，因此 Web 页数据在无线网络上的传输会带来较重的负担。(2)移动设备的微型浏览器屏幕尺寸相对 PC 机小很多。目前 Web 页大多是为 PC 机设计的，要在移动设备上浏览将需要大量的水平和垂直滚动操作，而在该类设备上实现这些操作往往比较困难，因此，在移动设备上浏览 Web 页信息具有相当的难度。鉴于移动用户往往更加关注网页的相关主题内容[1]，本文建立了一种基于WAP的移动搜索模型，该搜索引擎从检索到的结果页中抽取出与查询请求最为相关的主题内容，并使用WAP协议的标记语言格式(WML或xHTML)组织输出，使系统能够有效解决上述问题。

1 基于 WAP 的移动搜索模型

本文基于WAP的移动搜索引擎采取的是3段式的工作流程：网页搜索，网页预处理和内容服务。系统结构见图 1。

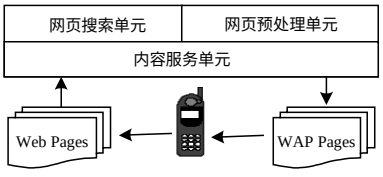

图 1 基于 WAP 的移动搜索模型

在图 1 中，网页搜索单元是移动搜索引擎的网页搜集单元，负责从 Internet 上搜索出符合查询请求的一系列网页。网页预处理单元对从搜索单元获取的 Web 页进行分割处理，并提取出网页中与查询请求最相关的主题内容。内容服务单元接收用户的查询请求进行检索，并输出搜索结果。

1.1 网页搜索单元

本文的网页搜索单元充分利用现有资源，利用现有成熟的、独立的搜索引擎(如 Google, Yahoo!等)实现对 Internet 上信息的搜索，不需要使用复杂的检索机制，也不需要建立和维护庞大的索引数据库，节省搜索成本。该单元的工作算法如下：

Begin
选择成员引擎。由用户根据个人的喜好或者以往的经验对系统给出的成员引擎进行选择；
查询请求转发。根据选择的成员引擎的搜索规则将用户提交的查询请求转发给成员引擎；
成员搜索。成员引擎利用自身的检索机制完成搜索，并返回一系列网页构成原始结果集。
End

1.2 网页预处理单元

网页预处理单元只处理原始结果集中当前依然存活的网页。该单元的工作算法如下：

Begin
判断网页文档类型
如果是 HTML 文档，则执行以下动作：
抽取出网页中与查询请求最为相关的主题块作为网页的结果内容块，并生成结果元组空间；
如果是 WAP 协议标记语言文档，则直接读取网页内容存放到结果元组空间；
End

抽取出网页中与查询请求最为相关的主题块，即在网页

基金项目：国家自然科学基金资助项目(60072006)
作者简介：陈 明(1949－)，男，教授、博士生导师，主研方向：分布式并行计算；孙丽丽，博士研究生
收稿日期：2007-03-25 **E-mail：**cuplily001@yahoo.com.cn

中寻找与查询请求相关度 *DoS* 最高的主题块。首先找出网页中的主题块，即能够反映出网页主要内容的区域。然后在这些主题块中按照一定的策略查找出与查询请求最为相关的主题块，具体步骤如下：

(1)分割网页成块

考虑到分割块的语义内聚性和分割粒度，可使用基于视觉的页面分割算法VIPS[2]进行分割。该算法建立语义树中每个节点均带有一个*DoC*值，表示节点的内容内聚度，*DoC*的值越大，表示节点的内容联系越紧密，反之，越松散。

(2)提取主题块

文献[3]中指出块的价值反映了块内容与 Web 页的主题内容的相关性。这里的“价值”并非注意力概念，注意力代表的是一个主观上的印象。当人们第一眼看到一个网页时，他们的注意力很可能会被页面中色彩鲜艳的图片或者生动的广告所吸引，通常这些内容并不是网页内有价值的部分。块的价值是一个客观的概念，它应该是从作者角度定义的，而非用户角度。

为了形式化地表示块的价值，文献[3]定义了一个块价值模型，将块的属性映射为一个权值。该模型将块的价值分成 3 个层次，其权值分别对应为 1, 2, 3，如表 1 所示。

表 1　块价值模型

级别	描述	权值
1	噪音信息，如广告，版权等	1
2	有用信息，但与网页的主题不相关，如导航栏、目录等；或者是与网页主题有关的信息，但不具备突出的重要性，如相关主题等	2
3	网页中最突出的部分，例如大字标题，主要内容等	3

块的属性包括空间属性和内容属性。空间属性反映的是块在页面中的位置和大小等，内容属性反映的是块包含的图片数目和大小、超级链接数目和文本长度等。具体参考文献[1, 3]。

然后利用神经网络或支持向量机等学习算法训练出一个模型，以通过块的属性评价出网页中各块的价值权值。本文利用文献[3]的块价值模型评价出网页中各语义块的价值权值。基于此，提取出其中权值为 3 的块。

(3)寻找与查询请求最为相关的主题块

一个网页可能存在多个主题内容，通过步骤(2)得到的层次为“3”的块可能存在多个。

由于移动设备小屏幕的限制，因此需要从中选择出与查询请求最为相关的块。这就需要通过某种策略衡量各个提取块与查询请求的相关度 *DoS*。

传统的搜索引擎通常是根据查询关键词与网页文档的索引词(即特征项)的匹配程度检索出匹配文档，而索引词权值的获取主要依赖于词在文档中出现的词频。因此，本文利用查询关键词在块中出现的词频计算相关度 *DoS*。

基本算法如下：

1)将查询请求短语转换成查询关键词。其中，如果查询请求短语表现为一中文的自然语句，则应用正向减字最大匹配法[4]分词；

2)去除中文查询关键词中“的”、“在”等“停用词”[4]；

3)记录步骤(2)后生成的所有关键词 k_i 构成集合 $K=\{k_i\}$, $i=1,2,\cdots,N$，N 为关键词的总数；

4)对于块 j，使用式(1)在集合 K 上计算 DoS_j

$$DoS_j=\frac{\sum_{i=1}^{N}KF_{ij}}{\sqrt{\sum_{j=1}^{M}\left(\sum_{i=1}^{N}KF_{ij}\right)^2}} \tag{1}$$

其中，N 为关键词的总数；M 为步骤(2)中提取的主题块的总数；KF_{ij} 为关键词 k_i 在块 j 中出现的词频；$DoS_j\in[0,1]$。根据式(1)，块 j 中关键词的词频越大，DoS_j 的值就越大，则块内容与查询请求的相关度就越高。

(4)生成结果元组空间

针对所有分析网页生成一个元组空间，该空间由四元组<*RelativityID, PageURL, PageTitle, ThemeContent*>组成。其中，*RelativityID* 表示网页与查询请求的相关度排序序号；*PageURL* 代表网页的访问地址；*PageTitle* 代表网页的标题，是网页的元数据，这 3 个元组的内容来自从成员搜索引擎返回的搜索结果列表；*ThemeContent* 存放最终输出给用户的内容，如果网页为 HTML 文档，则该元组为网页内与查询请求最相关的主题内容，即网页中 *DoS* 值最高的块内容；如果网页为 WAP 协议标记语言文档，则该元组为该文档的全部内容。

1.3　内容服务单元

根据分析统计，用户平均察看搜索引擎返回搜索结果不会超过 2 页[5](每页 10 个条目)，因此，为了提高系统响应速度，降低网络传输量，内容服务单元检索初始时只对原始结果集的前 20 个存活文档进行预处理，并返回这 20 个条目，条目信息包括标题和网页URL2 个要素。

内容服务单元采取的工作算法如下：

Begin

接收用户提交的查询请求；

调用搜索单元检索得到原始结果集 R；

读取 R 的前 20 个存活文档构成处理集 P；

调用预处理单元对文档进行处理，得到结果元组空间；

以列表形式返回匹配条目给用户，显示时采取分页方式，每页 5 个条目；

Do

分析用户行为

如果用户选择查看某匹配条目，则从元组空间中读取出该条目的 ThemeContent 以 WAP 标记组织输出给用户；

如果用户选择查看 R-P 文档的链接，则从 R 中依赖于原顺序继续读取 5 个存活条目添加到 P 中，并对这 5 个文档分别顺序执行 d,e,f 的动作；

如果用户选择查看 ThemeContent 中的超链，则执行 d 的动作，然后用 WAP 标记输出元组的 ThemeContent 给用户。

While(用户退出该查询请求的浏览)

End

2　模型实现与评价

系统实现时采用 Tomcat 作为 Web 服务器，使用 WML+JSP 作为 Web 页面脚本，利用 MySQL 作为后台数据库。图 2 给出了基于 WAP 的移动搜索引擎的体系结构。系统采用了 3 层体系结构：表示层，业务逻辑层和数据层。其中，表示层负责系统的用户界面显示，包括移动搜索接口和结果展示模块；业务逻辑层是系统的核心层，完成移动搜索引擎的控制处理流程，包括成员引擎搜索模块、预处理模块和显示处理模块，其中的显示处理模块负责以 WAP 标记语言格式组织返回给用户的结果内容；数据层包括资源规则库、词典

（下转第 209 页）