



## 个性化移动元搜索引擎研究与设计<sup>\*</sup>

何震苇<sup>1</sup>, 邹若晨<sup>2</sup>, 钟伟彬<sup>1</sup>, 严丽云<sup>1</sup>

(1. 中国电信股份有限公司广东研究院 广州 510630; 2. 华南理工大学 广州 510006)

**摘要:** 移动搜索已成为最重要的移动互联网应用之一, 智能终端技术和搜索引擎技术的发展为改进用户的搜索体验提供了可能。通过分析移动用户的个性化搜索需求, 提出了个性化移动搜索代理模型, 结合移动智能终端的能力和商业搜索引擎, 提供了符合用户个性化搜索习惯和多样化内容需求的移动搜索服务。

**关键词:** 移动搜索; 个性化; 元搜索; 搜索引擎

**doi:** 10.3969/j.issn.1000-0801.2013.05.008

## Research and Design on Personalized Mobile Search Agent

He Zhenwei<sup>1</sup>, Zou Ruochen<sup>2</sup>, Zhong Weibin<sup>1</sup>, Yan Liyun<sup>1</sup>

(1. Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China;

2. South China University of Technology, Guangzhou 510006, China)

**Abstract:** Mobile search has become one of the most important mobile internet applications. With the development of intelligent terminal technology and search engine technology, it's possibility to improve the user's search experience. The personalized search requirement of mobile users was analyzed, and a personalized mobile search agent model was proposed, in conjunction with the mobile intelligent terminal capacity and commercial search engine's ability to meet users' personalized search habits and diverse needs of mobile search service.

**Key words:** mobile search, personalized, meta search, search engine

### 1 引言

随着手机上网人数的快速增长和网络信息的飞速膨胀, 移动搜索已经成为人们利用移动终端进行信息检索的最主要工具, CNNIC 的最新报告显示, 移动搜索已成为国内仅次于手机即时通信的第二大移动互联网应用。移动搜索与传统 Web 搜索相比, 对信息精度的要求更高、对个性化搜索服务的需要更迫切, 因此, 如何提高移动搜索上的信息查询准确率、搜索结果返回的精度, 满足搜索用户的个性化需求成为亟需解决的问题。

较好地解决这些问题的根本思路就是在移动终端和搜索引擎之间引入搜索代理, 正确理解用户搜索意图, 返

回精确的搜索结果, 简化用户搜索操作。

### 2 个性化移动搜索的基本要求

移动搜索基本过程与 PC 搜索类似, 包括 3 个步骤: 提交搜索请求, 分析查询语义, 获取搜索结果。由于终端设备和用户群体的差异性, 移动搜索具有其显著特点, 如移动性, 用户可能会边走边搜, 而在移动中要输入搜索文字是很困难的; 多元化, 移动用户搜索的信息多种多样, 既可能是吃喝玩乐信息, 也可能是商品比价、健康医疗信息, 单一的搜索引擎往往难以覆盖用户的信息需求; 隐私性, 要挖掘个人兴趣必须搜集个人信息, 而用户的隐私又要得到合理保护。

<sup>\*</sup> “核心电子器件、高端通用芯片及基础软件产品” 国家科技重大专项基金资助项目“移动智能终端操作系统研发”(No.2012ZX01039002-001)



因此,一个完善的移动个性化搜索引擎应重点考虑以下几个方面。

- 使用便捷:不仅要提供用户查询词的自动补全和相关搜索提示等功能,还需要提供语音、菜单等更高效的搜索请求输入方式。
- 个性化服务:所谓个性化,就是能够感知人及人所在的环境,针对“此时、此地、此人”建立模型,正确理解用户的搜索意图。
- 保护隐私:允许用户定义隐私基准,控制个人信息开放的范围。
- 响应迅速:迅速响应用户的搜索请求,与传统搜索引擎相比没有明显时延。

### 3 移动个性化搜索代理相关技术

作为一种新型搜索技术,个性化移动搜索的研究仍处于起步阶段。这种新兴的搜索是搜索技术在移动平台上的延伸,真正打破了地域、网络和硬件的局限性,满足了用户随时、随地的搜索需求。根据第2节提出的移动个性化搜索代理的需求,对其中较为关键的实现技术进行了细致分析。

#### 3.1 元搜索引擎

元搜索引擎,是指在统一的用户查询界面与信息反馈形式下,共享多个搜索引擎的资源库为用户提供统一的搜索服务。元搜索引擎是对搜索引擎进行搜索的引擎,元搜索引擎与一般搜索引擎的最大不同在于,它可以没有自己的资源库和机器人,而只是充当一个中间代理的角色,接受用户的查询请求,将请求翻译成相应搜索引擎的查询语法。在向各个搜索引擎发送查询请求并获得反馈之后,首先进行综合相关度排序,然后将整理抽取之后的查询结果返回给用户。

因为元搜索引擎的结果是基于独立搜索引擎的,所以在功能上,这些排列顺序会根据网站权重、网站关联度等进行改变,实现上将会受到独立搜索引擎的限制。元搜索引擎的覆盖比较广,但是检索的准确率却不容易控制,而个性化结果整合排序就显得尤为重要。

#### 3.2 用户兴趣建模

个性化搜索的核心是根据用户的行为,建立一套准确的个人兴趣模型。用户兴趣即用户的检索习惯,针对用户的搜索习惯进行信息检索,可以极大地提高信息传递的准确性。对于移动搜索代理来说,不仅是移动搜索的内容,移

动设备所提供的其他信息(如地理位置)也同样重要。个性化移动搜索系统的核心内容便是通过数据挖掘技术在用户数据信息中自动地发现有用的信息,构建用户兴趣模型。

用户建模主要包括如下两个方面。

##### (1) 兴趣模型结构的构建

用户的兴趣可以表现为不同的兴趣类别,各个兴趣类别之间在结构或者语义上存在一定的联系,模型构建的意义在于发现并组织这些联系,确定以何种数据模型表达用户的兴趣,使用户兴趣模型清晰、高效地为个性化信息检索服务。

##### (2) 兴趣类别以及兴趣度权值的计算

用户的兴趣类别千差万别,对不同类别表现的兴趣也有很大差异,用户兴趣建模用兴趣类别和兴趣度权值体现这种差异,在得到兴趣模型结构的基础上,计算不同的兴趣类别,并对不同的兴趣进行分析,以区别用户对不同类别喜好的程度高低。现有的用户兴趣模型中,使用较多的主要有空间向量模型、GauchS的基本体的表示方式以及概念层次模型。

图1是一种比较理想化的用户建模方式,即通过终端全面收集与用户相关的信息源,包括用户搜索历史及点击记录、用户浏览过的网页、用户位置信息、用户所收藏的信息以及用户发布的信息等内容。

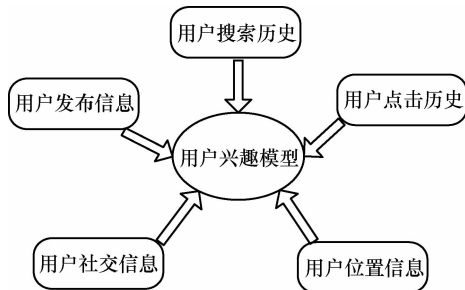


图1 理想化的用户兴趣模型

在用户兴趣建模中,最常用的方式是将显示和隐示两种方式结合起来,通过显示的方式获取静态用户信息,通过隐示的方式获取动态用户信息。具体来说,用户兴趣挖掘信息的主要来源有以下几个方面:

- 用户输入搜索引擎的关键词;
- 用户的浏览历史,即用户在搜索结果中点击浏览的页面,通过计算用户浏览页面与兴趣分类模型的相关度,可以识别用户的浏览兴趣;
- 用户位置,可通过用户的位置坐标识别用户的位置属性,如街道、城区、城市、省份等;

- 用户社交信息, 用户所属的社交网络群组的兴趣特征;
- 用户发布的信息, 如用户自定义的标签。

作为建模数据的用户信息需要经过谨慎选择, 因为用户信息对用户兴趣模型的性能有非常重要的影响, 因此, 建模数据只能是那些真正反映用户兴趣的信息, 否则构建出来的用户兴趣模型的性能将得不到保证。

### 3.3 意图分类

用户发出的每个搜索请求背后都隐含着潜在的搜索意图, 如果能够通过查询词汇, 自动找出背后的用户搜索意图, 然后针对不同的搜索意图, 提供不同的检索方法, 将更符合用户意图的搜索结果排在前列, 无疑会增加搜索引擎用户的搜索体验。例如, 用户搜索“广州 天气”的时候, 会主动将当天的气温等情况列在搜索结果最前面。

搜索意图分类的任务, 就是确定用户输入的查询属于哪个预定义的意图类别。多数分类技术是根据输入数据集建立分类模型的系统方法。其通过学习算法确定分类模型, 该模型不仅要很好地拟合输入数据, 还要能够正确地预测未知样本的类别。因此, 学习算法的主要目标就是建立具有很好的泛化能力的模型, 即建立能够准确地预测未知样本类别的模型。

图2展示了解决分类问题的一般方法。首先, 需要一个训练集, 它由类标号已知的记录组成。在使用训练集建立分类模型前, 先要对数据进行预处理, 以帮助提高分类的准确性、效率和可扩展性。预处理包含大量以复杂方式相关联的不同策略和技术, 例如聚集、抽样、维规约、特征子集选择、特征创建、离散化和二元化、变量变换等。这些项目分为两类, 即选择分析所需要的数据对象和属性以及创建/改变属性。例如记录用户提交搜索的星期数, 就可能与搜索类型分类无关; 此外有些属性也可能是冗余的。因此需要对数据进行维规约以及特征子集选择, 以帮助其在

学习阶段就消除无关或冗余的属性。

其次, 预处理从原始的、大量的特征项中筛选出区分能力最强、最有代表性的特征项。数据经过预处理之后, 将会转换成分类算法所能接受的数据类型。此时分类算法已从训练集的学习过程中得到了一个目标函数, 也称作分类模型, 该模型将运用于对未知类别的用户查询进行分类。

### 3.4 语音识别

语音识别能够代替繁琐的键盘或手写输入, 大大提升了移动搜索的便捷性。目前 Android、Windows Phone 等主流的智能终端均提供了语音识别 SDK, 这些 SDK 封装了服务端语音识别引擎的 API, 为终端应用提供 Speech2Text 的转换能力。

以 Android 的语音识别 SDK 为例, 整个语音识别过程包括以下步骤。

(1) 判断当前设备是否支持语音识别能力, 只有终端安装 Android 语音识别软件并开启网络连接时, 才能进行语音识别。

(2) 开启语音识别功能, 启动语音识别界面, 提示用户输入语音。

(3) 用户进行语音输入时, 通过明显的语音停顿或点击“完毕”按钮完成语音输入。

(4) 语音识别 SDK 捕捉用户输入的语音信号, 发送给谷歌语音识别引擎进行处理。

(5) 谷歌语音识别引擎根据语音信号的声学特征、语言模型及发声词典, 寻找能够以最大概率输出该语音信号的词串, 返回给终端。

语音识别 SDK 将语音识别引擎返回的词串通过回调函数返回给终端应用。

## 4 个性化移动搜索代理设计

### 4.1 架构设计

个性化移动搜索代理采用 C/S 架构, 如图3所示, 移动用户通过客户端输入搜索请求, 客户端将用户的搜索请求和用户个性化标签通过 HTTP+JSON 协议发送给搜索代理服务器, 代理服务器将用户的搜索请求转发给百度、谷歌等商业搜索引擎, 并将搜索引擎返回的搜索结果进行处理后返回给用户。代理服务器提供数据加工和页面重排服务; 而客户端发挥着收集、分析和提取用户资料的作用。出于对用户个人隐私保护的考虑, 搜索过程中一些简单的任务, 如整理用户浏览记录、更新点击率和搜索记录、创造用

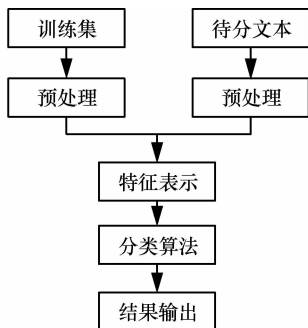


图2 一般分类过程

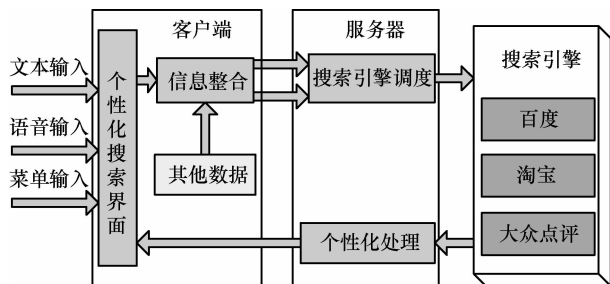


图3 搜索代理系统架构

户特征等都在客户端处理。

客户端不仅能够记录用户的搜索历史、搜索位置,还能利用智能终端的传感器搜集与用户搜索行为相关的温度、速度等关联的场景信息,整合成完整的用户个性化模型。用户的查询信息在客户端与用户喜好标签关联,以体现用户的搜索意图,客户端还负责对自服务器的搜索结果数据进行个性化展示。

代理服务器接收客户端的请求,根据用户的搜索意图调用合适的成员搜索引擎,将搜索引擎的结果进行个性化排序后返回给客户端。

#### 4.2 客户端设计

客户端采用基于 PhoneGap 的混合应用架构,如图4所示。在查询过程中,用户的输入方式有很多种,除了传统的文本输入方式外,还可以使用语音输入、菜单输入方式。客户端会依据用户的兴趣模型优化用户的查询请求,缩小查询范围,例如在查询餐馆时增加城市关键字。同时客户端会从用户兴趣模型中提取出与本次查询相关的用户标

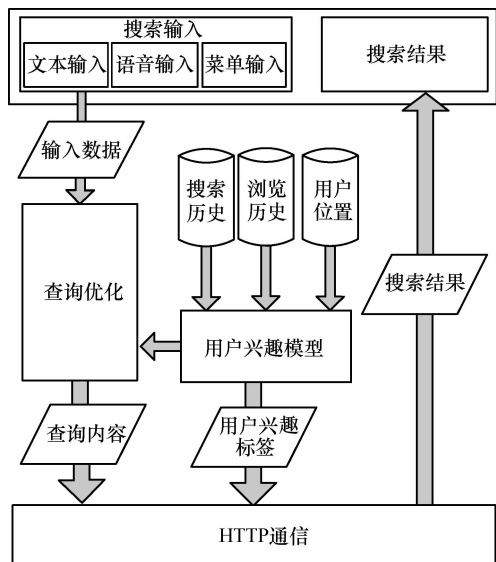


图4 客户端结构

签,如查询餐馆时,打上川菜、粤菜等标签。查询的关键字和用户标签以JSON格式封装,通过HTTP发送给代理服务器。代理服务器返回的查询结果也以JSON格式封装,客户端通过Web View框架进行解释和呈现。

客户端还记录用户的搜索关键字、搜索位置、搜索结果页面的浏览历史,为了保护用户隐私,这些详细日志信息只保存在客户端数据库中,不发送给服务器,只有搜索的关键字和与搜索会话相关的用户标签才会发送给服务器。用户还可以通过设置位置精度(街道、城区、城市、省份等)和内容精度(即用户标签级别)来保护个人隐私。

#### 4.3 服务端设计

代理服务器的架构如图5所示,代理服务器负责根据用户的搜索请求选择成员搜索引擎,向成员搜索引擎转发搜索请求,按用户喜好整合搜索结果并以JSON格式返回给客户端。

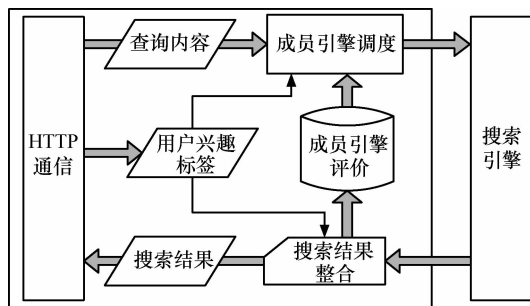


图5 代理服务端结构

##### (1)成员引擎调度

成员搜索引擎的选择要根据用户的搜索意图和成员引擎的评价进行综合判断,用户的搜索意图可通过搜索关键字和用户标签判断,比如用户搜索商品信息就选择淘宝引擎,搜索餐馆信息就选择大众点评引擎。成员引擎的响应速度、结果质量也是成员引擎选择的重要依据。

##### (2)搜索请求转发

不同搜索引擎的搜索条件可能存在差异,将原始的搜索请求转换为成员搜索引擎的目标搜索条件有助于提升搜索的精度。例如用户输入“1 000 元手机”,映射到淘宝引擎的搜索输入条件即为“商品:手机+价格:800~1 200”。

##### (3)搜索结果整合

根据用户的兴趣计算搜索结果和用户的关键词和标签的相关度,然后考虑成员搜索引擎的权重、搜索结果的重复度等因素,最后计算搜索结果的权重值,对结果进行整合排序。整合后的搜索结果封装成JSON对象,通过

HTTP 返回给客户端。

为了提高代理服务器的性能,可以引入负载均衡器构建代理服务器集群,并通过分布式缓存技术保存频繁访问的搜索结果。

## 5 结束语

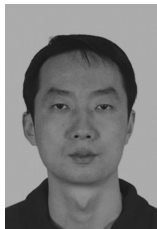
个性化服务是移动搜索的一种趋势,也是一个研究热点,为了满足移动搜索的个性化需求,本文研究了个性化技术和元搜索技术,结合智能终端能力,设计出了个性化移动搜索代理的基本架构,能够在一定程度上提升移动搜索的效率和准确性。今后将持续改进用户兴趣模型,优化服务器性能,使之能在用户检索效率和用户体验方面得到更好的提高。

## 参考文献

- 1 Kenneth Wai-Ting Leung, DikLun Lee, Wang-Chien Lee. PMSE: a personalized mobile search engine. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4)
- 2 许天亮,王义峰,曾平. 个性化元搜索引擎技术研究. 电子科技, 2008(1):56~59

(上接第 44 页)

### [作者简介]

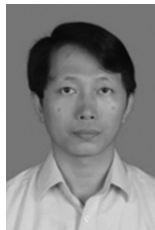


陆钢,男,硕士,中国电信股份有限公司广东研究院高级工程师,主要从事移动互联网、云计算技术研发工作。

区洪辉,男,硕士,中国电信股份有限公司广东分公司高级工程师,主要从事电信业务平台、云计算技术、移动互联网技术规划与实施工作。

- 3 王忠,程磊. 基于元搜索引擎的个性化 Web 信息采集. 计算机工程与设计, 2009(7):3117~3119

### [作者简介]



何震苇,男,硕士,中国电信股份有限公司广东研究院工程师,主要从事终端应用和云计算平台研发工作。

邹若晨,男,华南理工大学硕士研究生,主要研究方向为移动互联网应用和平台研发。

钟伟彬,男,硕士,中国电信股份有限公司广东研究院工程师,主要从事云计算相关技术研发工作。

严丽云,女,硕士,中国电信股份有限公司广东研究院工程师,主要从事移动互联网、云计算及虚拟化技术研发工作。

(收稿日期:2013-04-15)

梁柏青,男,学士,中国电信股份有限公司广东研究院部门技术总监、高级工程师,“核心电子器件、高端通用芯片及基础软件产品”国家科技重大专项基金资助项目“移动智能终端操作系统研发”的负责人,主要从事电信网络技术研究,交换设备、综合数据网络产品和电信终端的设计和研发工作。

李蓉蓉,女,学士,中国电信股份有限公司广东研究院高级工程师,主要从事平台相关技术研究和研发工作。

(收稿日期:2013-04-15)

\*\*\*\*\*

### · 简讯 ·

## 2013 年度中国通信学会科学技术奖励工作会议在京举行

2013 年 4 月 24 日,中国通信学会在工业和信息化部电信研究院 3G 大楼报告厅举行“2013 年度中国通信学会科学技术奖励工作会议”。出席本次工作会议的领导有:国家科技奖励办公室副主任陈志敏、工业和信息化部科技司副司长韩俊、中国科协学会学术部副部长兼学会管理处处长范唯、工业和信息化部电信研究院副院长于生多。中国通信学会副理事长兼秘书长张新生、中国通信学会副秘书长秘建虎、副秘书长张英海出席大会,并为获得 2012 年度中国通信学会科学技术奖的 35 个获奖项目代表颁奖。