# 移动 Web 搜索关键技术研究

张金增

## 1 引言

随着移动通信和 Internet 在人们日常生活中的日益普及，移动通信带宽的大幅度提高和移动终端功能的逐渐增强，传统的服务已经不能满足用户多元化的需求，人们希望随时随地利用移动终端访问互联网上的服务，从中获取丰富的信息。移动互联网实现了 Web 和移动通信的逐步融合，使其成为产业界备受关注的领域。

随着 3G 时代的到来，越来越多的用户使用移动终端能够便捷地访问网络，根据中国互联网络信息中心发布的最新报告显示，手机网民数已达 1.137 亿，并呈直线上升趋势，并且使用手机上网的用户会越来越多。而信息搜索是用户在访问网络时最经常进行的活动之一。在日常生活中，人们经常会碰到很多“now and hereò的问题，需要查询与其正在进行的活动相关的信息：

- 我的朋友现在位于什么位置？
- 电影院现在正在放映什么电影，离自己位置最近的影院有无剩余的票？
- 附近有没有比较好的吃晚餐的地方？现在有一些什么优惠活动？
- 我需要停车，现在离我最近的有停车位的停车场在哪里？

由以上问题可以看出，人们使用移动设备搜索时大多数需求都与位置密切相关，但使用传统的搜索引擎仅仅利用纯粹的文本关键字搜索，用户往往不能获得理想的查询结果。此外，与传统互联网搜索环境相比，移动终端受到了屏幕尺寸小、网络带宽有限等限制。这些不同点为新环境下移动 Web 搜索带来了许多新的挑战。

因此，本研究在移动环境下，根据移动用户的需求，将地理数据与 Web 数据进行无缝的集成；在此基础上进行高效的面向移动用户的查询处理，获得高度精确的满足用户需求的结果，从而为用户提供“Near by Now”的服务，具有非常重要研究价值。为 3G 时代下的移动 Web 搜索提供了一种新思路，具有十分广阔的应用前景。

## 2 移动 **Web** 搜索概述

与互联网搜索和移动数据库查询技术相比，移动 Web 搜索具有独特的特点。本部分结合相关研究工作，对移动 Web 及其特性进行了分析和总结。在此基础上，指出了移动 Web 搜索与互联网搜索存在的差异。

### **2.1** 移动 **Web** 基本概念及其特点

移动互联网是指移动用户从自身实际需求出发，能够通过无线终端随时随地的通过无线方式接入互联网。传统的移动互联网是一个封闭的网络，其封闭性体现在网络、终端和应用三个方面。封闭的特性制约了移动互联网的发展。新型移动互联网具备如下特点：

开放性：开放性体现在网络开放、应用接口开放、内容和服务开放等多个方面，用户拥

有选择的权利。

分享和协作性：在开放的网络环境中，用户可以通过多种方式与他人共享各类资源，可以实现活动参与，协同工作。

创新性：结合 Web2.0 与移动网特征，移动互联网能够为用户提供无穷无尽的创新性业务。

开放、分享和创新构成了移动互联网的核心特征。随着移动互联网的深入发展，移动互联网现有的垄断性、封闭性终将被打破，开放性将成为移动互联网服务的基本标准，用户将具有更大的自主性和更多的选择，用户角色由被动的信息接受者转变成为主动的内容创造者，移动终端的智能性将进一步增强，用户之间的通信和内容体验将更具有交互性。

### 2.2 移动 Web 搜索与互联网搜索的异同点

移动 Web 搜索是指以移动网络为数据传输承载，将分布在传统互联网和移动互联网上的数据信息进行搜集整理，供手机用户查询的业务。移动搜索作为搜索技术与移动通信技术的一种结合体，融合了两种技术的各自特点。移动搜索的出现，真正打破了地域、网络的局限性，满足了用户随时随地的搜索服务请求。

庞大的手机用户群成为移动搜索的潜在用户，该类用户区别于互联网用户的特征以及移动网的特点，对搜索技术的功能实现提高了更高的要求。移动搜索与互联网搜索存在本质区别，主要表现在搜索方式、搜索要求、搜索渠道和搜索内容等多个方面。详见表 1

| | 移动 Web 搜索 | 互联网搜索 |
|---|---|---|
| 终端特点 | 屏幕较小、功能单一、普及率高、体积小、携带方便、承载网络覆盖面面大 | 大屏幕、功能丰富、普及率较低、体积较大、携带不便、承载网覆盖面小 |
| 搜索方式 | 关键字搜索、自然语句搜索 | 目录检索、关键字搜索 |
| 搜索需求 | 准确性、便捷性、个性化 | 准确性、海量性、快速性 |
| 搜索渠道 | 短信、搜索门户、搜索栏、IVR | 搜索门户、搜索栏、浏览器地址栏 |
| 搜索内容 | Wap 网站内容、传统互联网内容、运营商及服务提供商内容、传统信息提供商及黄页内容 | 以互联网网站内容为主，信息量十分丰富 |
| 搜索目的 | 搜索需要的内容、定制需要的服务 | 搜索需要的内容和站点 |
| 搜索限制 | 无 | 存在网络接入限制 |
| 搜索费用 | 流量费、服务定制费、部分搜索服务需要单独付费 | 免费 |

表 **1**　移动 **Web** 搜索与互联网搜索的差异

## 3 移动 Web 搜索基本框架及关键性技术

作为一种新型搜索技术,移动 Web 搜索的研究仍处于起步阶段. 这种新兴的搜索是传统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。为了满足这些需求，需要提出一系列对移动数据进行表示、模型构建、索引和信息检索的新技术。这一节首先提出移动 Web 搜索框架，然后从地理标记 Web 资源、混合索引的构建、面向移动用户的查询处理、查询结果的排序与可视化几个方面，对移动搜索的关键性技术进行分析。

### 3.1 移动 Web 搜索基本框架

移动环境其位置动态变化，屏幕狭小、计算资源有限等特点对传统的文本搜索提出了高精准的查询要求，给移动Web搜索带来了许多新的挑战。在移动Web搜索领域存在着许多研究问题: 地理标记Web资源、集成更新、混合索引的构建、面向移动用户的查询处理、查询结果的排序与可视化等. 有些问题已经得到了一定程度的研究,而有些问题还处在研究的初步阶段。为了给出一个全面的认识，我们提出了移动Web搜索的整体框架，如图2所示。该框架被划分为四个模块：

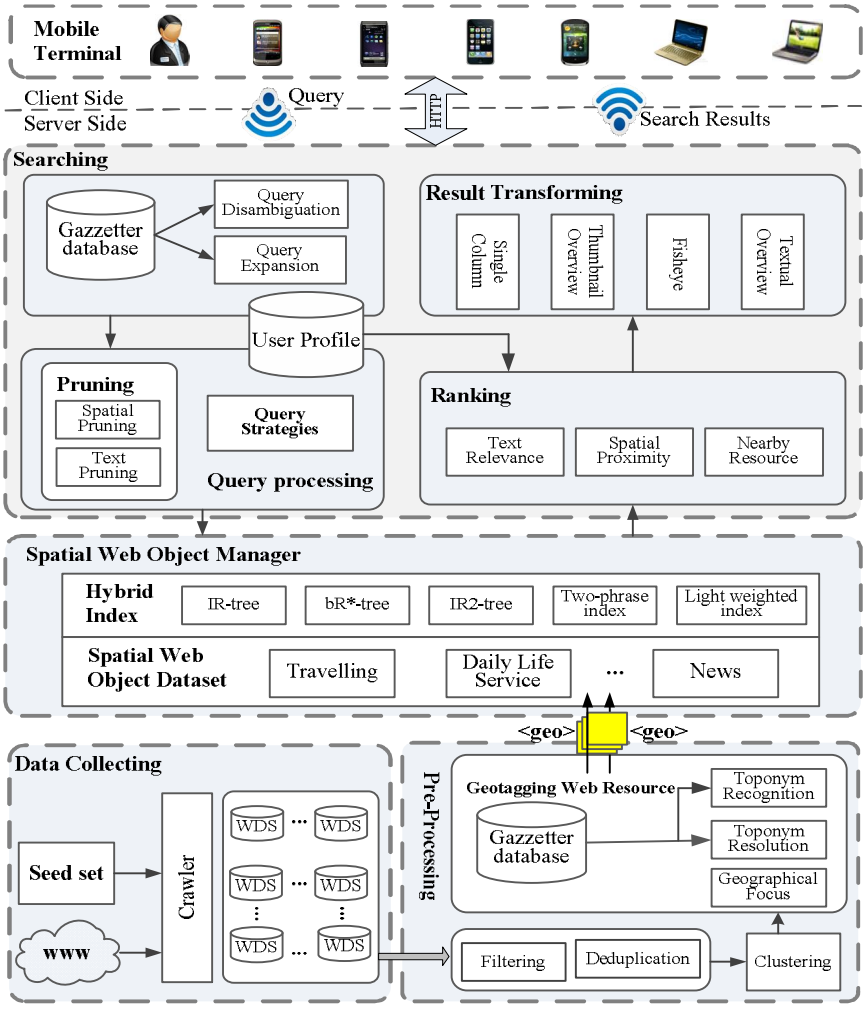


图1 移动Web搜索系统框架

- 数据搜集模块(Data Collecting)：通过种子节点集合，爬虫从 WWW 发现和下载 Web 页面，接收 URL，下载页面，分析内容，并沿着出链接重复执行一系列操作，从而获得初始的粗糙数据集。
- 预处理模块(Pre-collecting)：数据搜集阶段完成后，需要对 Web 数据源进行清洗和去重操作，并对过滤后的数据按内容所在的领域进行聚类。为了支持移动搜索，需要标记出 Web 资源所对应的地理位置或者覆盖的地理范围，完成这项任务需要地名识别、地名分辨和覆盖地理范围的确定三个步骤，在此基础上，就可以获得空间 Web 对象数据集。
- 索引模块(Indexing)：对空间 Web 对象构建索引需要综合考虑地理空间和文本两个方面，为了提高搜索的效率和访问的准确率，需要对其构建混合索引，将空间索引和文本索引进行无缝的集成。混合索引的构建用来检索出与用户需求高度相关的信息。
- 搜索模块（Searching)：用户通过移动终端设备提交查询，需要对提交的查询消除歧义并扩展；然后进行查询处理，对返回的结果按照文本相关性、距离相近性及周围环境进行综合排序；由于移动设备固有的一些固有局限，需要对用户查看的结果页面进行一定的转换处理；最终将搜索结果返回给移动终端。

### 3.2 移动 Web 搜索关键性技术及存在的挑战

移动环境其位置动态变化，计算资源有限等特点给移动 Web 搜索带来了许多新的挑战，本文的研究内容包括地理标记 Web 资源、为空间数据和文本信息所组成的空间 Web 对象构建混合索引、基于位置的查询处理和查询结果的处理等关键性问题。

**(1)** 地理标记 **Web** 资源

许多 Web 资源像商业、新闻等 Web 页面都包含大量与位置相关的信息，再加上地理位置对移动搜索结果的精确性具有决定性的作用。因此，如何准确有效的找出 Web 资源对应的地理位置是一个关键性的问题。

对于给定的 Web 资源，准确的标记出所对应的地理位置或覆盖的地理范围大致需要三个步骤：地名识别（toponym recognition）、地名分辨（toponym resolution）和覆盖地理范围的确定 （Geographical focus）。

(a) 地名的识别：需要处理 geo/non-geo 歧义的问题，目前最普遍采用的是自然语言处理方法。但这种方法仅仅适用于分析静态的文档库，对于动态和不断更新的文档库却并不适合。为了解决该问题，可以采取混合的地名识别技术，使用词性标记（POS）和命名实体识别（NER）对位置短语进行标记，搜集这些地名采用基于规则的启发式方法和基于统计的处理相结合的方法。

(b) 地名分辨：主要任务是解决 geo/geo 歧义。最常用的方法就是为识别出的地名分配一个衡量其流行度的缺省值，并结合启发式规则来完成地名的分辨。但这种方法对于那些不是很出名的地方却无能为力。对于这个问题，一个初步的想法是可以采用分层地理本体的概念去解决，在分层地理本体中建立“分辨上下文”。

(c) 覆盖地理范围的确定（Geographical focus）：就是找出文档覆盖的地理区域。使用

层次本体去决定覆盖地理范围，在该层次结构中每一个分辨出来的地名都为它的父节点贡献分值，然后选择分值最高的本体节点作为该文档的地理聚焦点。另外一种普遍使用的策略就是选择出现频度最高的地名作为地理聚焦点。

**(2)** 混合索引技术

移动搜索需要检索与地理上下文相关的文档，这种需求要求索引建立以文本和位置为基础。R-tree、四分树、网格等是空间索引方法，而倒排索引、位图索引、签名文件是文本信息检索中有效的索引方法。这两类索引在针对不同的应用在各自的领域都得到了很好的发展。解决移动Web搜索的最简单的索引方法是先使用倒文本索引找出匹配关键字的结果，然后采用空间索引搜索进行空间搜索。但这种两阶段的索引方法对于提前决定需要获得top-k个结果，在第一阶段所需要的查询结果个数是非常困难的，并且CPU代价和I/O代价也非常高，所以这种方法不适合移动Web搜索。因此需要设计出一种综合考虑文本和空间位置的索引结构，使其有效地整合空间索引和文本索引以保证达到最优的搜索效果。

一种就是将用于文本检索的倒排文件和用于空间搜索的 R-tree 结合起来，使用倒排文件对 R-tree 进行扩充。新建立的 R-tree 的每个节点包含的信息是以该节点为根的子树的所有对象的位置信息和文本内容的概括。叶子节点指向对象的文本文档的索引文件，而非叶子节点的文档是将该节点所有孩子节点的所有文档的并集。在使用该索引结构进行搜索时，如何针对不同的应用进行设计，使其所占用的空间较小又能拥有高效的搜索效率，还需要进行深入的研究。

**(3)** 面向移动用户的查询处理

查询处理算法利用构建的混合索引方法去评估空间相近性和文本相关性。对于移动用户提交的查询，返回的结果与移动用户当前的位置密切相关，提交相同的查询，其时间、位置不同，得到的结果会有很大的差异，查询的结果是需要按照空间的相近性和文本的相关性进行排序。移动环境下最常见的几种空间查询 KNN 查询、RkNN 查询、Range 查询、Closest-Pair 查询等都需要采取不用的查询策略进行处理（比如深度优先、最佳优先等）。使用这些方法对空间 Web 对象进行搜索时，针对不同的场景，需要找出合适的处理策略，并找出约束条件，使得空间剪枝和文本剪枝在查询处理的过程中同时进行，从而有效的加速搜索的进程，对整个搜索具有极其重要的作用。

**(4)** 查询结果的处理

移动设备由于自身的特点只能为用户提供较小的显示区域，无法浏览大量的信息，如果用户被淹没于大量查询结果中，会导致用户的满意度下降。因此需要对查询结果进行优化处理，把用户最满意的查询结果以最简洁的方式按照某种顺序进行展示。

(a) 查询结果的排序：对于从空间Web对象数据库返回的大量结果，需要将与用户查询最相关的记录排列在靠前的位置，以提高用户的满意程度。对查询结果的排序需要综合考虑多个因素对排序的影响。不仅要考虑文本的相关性，还要考虑位置的相近性和周围环境对查询结果的影响，。

(b) 查询结果记录摘要的生成：通常一个查询结果包括许多数据项，但移动设备因其自身特点只能提供较小的显示区域，如果将查询结果的全部信息都显示，会大大降低用户的浏览效率。在实际中并不需要查询一个结果的所有信息，因此需要根据用户的查询、当前的位置等信息来选择合适的数据项，进一步将所过滤出的数据项进行文本摘要处理。

## 4 结论

随着移动网络的日益普及,用户的规模呈指数级增长，移动 Web 搜索越来越成为学术界和工业界共同关注的热门话题。本文提出移动 Web 搜索框架，然后从地理标记 Web 资源、混合索引的构建、面向移动用户的查询处理、查询结果的排序与可视化几个方面，对移动搜索的关键性技术进行分析，指出仍然存在的挑战和可能的解决办法。目前，虽然在移动环境下的 Web 搜索的某些方面提出了一些解决方法，但总体上缺乏系统性。总的来说对移动 Web 搜索研究仍然处于刚刚起步的阶段,仍有大量关键问题需要进行深入细致的研究。因此，移动 Web 搜索具有非常重要的研究价值和广阔的应用前景。