移动应用集成:框架、技术与挑战

马友忠[1)] 孟小峰[1)] 姜大昕[2)]

[1)](中国人民大学信息学院 北京 100872)

[2)](微软亚洲研究院 北京 100080)

摘 要 随着移动互联网的广泛普及和飞速发展,出现了大量的移动应用,其种类和数量还在不断增加.手机制造商、电信运营商和互联网服务提供商等纷纷推出自己的移动应用商店,移动应用已经成为互联网发展的一种新模式.移动应用的相关信息分布在应用商店、专业论坛及社交网络中,由于其信息的多样性、异构性、动态性,给移动应用集成带来了巨大挑战.移动应用集成的主要任务是研究如何把海量的移动应用及其相关信息有效地集成起来,为用户提供高质量的搜索、发现和推荐服务.移动应用集成还是一个比较新的研究领域,文中提出了一个移动应用集成的基本框架,对移动应用集成中的关键技术进行了分析总结,在此基础上对未来的研究方向及挑战进行了阐述.

关键词 移动应用;短文本;功能建模;移动应用集成;数据抽取;移动互联网;移动应用搜索

中图法分类号 TP311 **DOI 号** 10.3724/SP.J.1016.2013.01375

Mobile Application Integration: Framework, Techniques and Challenges

MA You-Zhong[1)] MENG Xiao-Feng[1)] JIANG Da-Xin[2)]

[1)](*School of Information, Renmin University of China, Beijing* 100872)

[2)](*Microsoft Research Asia, Beijing* 100080)

Abstract With the rapid development of the mobile Internet, large amount of mobile applications emerge, the types and number are still increasing. Handset manufacturers, carriers and Internet service providers also lunched their own application stores, mobile application has become a new model of Internet. The related information of the mobile applications exists in the App stores, professional forums and social networks; they are always diverse, heterogeneous and dynamic, so it is a very challenging task to integrate the variety of the mobile applications. The main task of the mobile application integration is to integrate the basic information of the applications and other related information efficiently, and to provide effective application search, application discovery and recommendation services. Mobile application integration is still a relatively new field of study, a framework of the mobile application integration is proposed in this paper, some related works on some key issues are analyzed, and finally several future research works are explained.

Keywords mobile application; short text; functionality modeling; mobile application integration; data extraction; mobile Internet; mobile application search

---

收稿日期:2011-12-24;最终修改稿收到日期:2013-05-09.本课题得到国家自然科学基金(61070055,91024032,91124001)、国家"八六三"高技术研究发展计划项目基金(2012AA010701,2012AA011001,2013AA013204)、中国人民大学科学研究基金(11XNL010)资助.

马友忠,男,1981 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为 Web 数据管理、云数据管理.E-mail: ma_youzhong@163.com. **孟小峰**,男,1964 年生,教授,博士生导师,主要研究领域为网络数据管理、云数据管理、移动数据管理、XML 数据管理、闪存数据库系统以及隐私保护. **姜大昕**,男,1975 年生,博士,微软亚洲研究院研究员,主要研究领域为数据挖掘和信息检索.

1 引 言

随着智能手机和其他移动设备的普及,移动互联网快速发展,海量的移动应用(Mobile Application,App)成了移动互联网的主要入口.根据瑞士信贷集团估计,到 2016 年,全球将会有 100 亿部联网的移动设备,智能手机的网络流量将会是今天的50 倍,而更多的移动设备也意味着更多的移动应用.苹果公司于 2008 年 7 月首次推出移动应用商店 App Store,获得了巨大成功.2012 年 10 月,应用数量已经超过 70 万,至 2013 年 5 月,官方应用商店 App Store 的应用下载量即将突破 500 亿次.同时,全球移动应用规模也在急剧扩大,手机制造商、电信运营商和互联网服务提供商等纷纷推出自己的移动应用商店,移动应用已经成为移动互联网发展的一种新模式.表 1 给出了几个比较有代表性的移动应用商店的基本情况.预计 6 月份,Google Play 商店的应用数量将超过 100 万.目前全球移动应用数量的规模在百万级别,与现有的 Web 网站和 Web 网页数量规模相比虽然还比较小,但是其现在的数量规模已经与 2000 年左右的网站和网页数量规模相当,并且还在不断增加之中.长尾理论提出者、《连线》的 Chris Anderson 曾提出“Web 已死,互联网万岁”,表示随着 iPhone/iPad 日渐成为主流计算终端,人们越来越习惯于通过移动应用软件获取信息.移动应用将逐渐超过浏览器,成为移动互联网的主要入口.

表 1 移动应用规模

公司	应用商店	应用数量	发布时间	类型
苹果	App Store	700000+	2008-07	手机制造商
谷歌	Google Play	800000+	2008-10	互联网服务提供商
黑莓	App World	70000+	2009-04	手机制造商
诺基亚	Ovi Store	100000+	2009-05	手机制造商
中国移动	Mobile Market	95000+	2009-08	电信运营商
中国电信	天翼空间	150000+	2010-03	电信运营商
微软	WP7 Marketplace	100000+	2010-10	互联网服务提供商
中国联通	Wo Store	30000+	2010-11	电信运营商

面对数百万的移动应用(未来还将继续增加),用户正面临着一个日益严重的挑战:如何才能快速找到自己想要的、适合自己的应用?而众多的移动应用开发者也面临着一个问题:如何把自己开发的应用推荐给用户?用户与应用开发者之间的供需矛盾日益突出.目前解决这一矛盾的方法主要有 3 种:

(1)移动应用商店.在移动应用发展的早期,移动应用主要出现在应用商店中,如表 1 中所列出的几个主要的应用商店.为了便于用户浏览、查找自己所需的移动应用,各应用商店都对数据进行了一些处理,包括分类、添加标签等.但是通过分析发现,目前的分类粒度比较粗,一般包含两个层次类别,大类数量在 20 个左右,由于应用的总体数量比较大,所以单个类别下的移动应用仍然比较多,用户要想快速定位到自己需要的移动应用依然很困难;另外不同应用商店的分类方式及类别名称不统一,各商店之间应用类别名称仅有 50%左右是一致的;各移动应用商店所提供的搜索功能大都是基于关键字匹配的简单搜索,搜索结果比较差,无法满足用户需求.

(2)第三方移动应用集成.为了解决移动应用商店存在的问题,出现了第三方移动应用集成服务提供商,其主要工作方式是从不同的应用商店中抓取移动应用信息,并对抓取到的应用信息进行进一步的处理,如重新分类、去重、添加标签等,在此基础上提供应用浏览、搜索功能.

(3)移动应用搜索与推荐.移动应用搜索与推荐是帮助用户快速找到自己所需应用的一种有效途径,目前已经有一些相应的解决方案.腾讯于 2012 年 6 月发布了海纳应用搜索,这是一款基于移动应用功能属性搜索的引擎.据腾讯介绍,海纳应用搜索是专门为用户提供移动应用搜索服务的智能搜索引擎,专注于 App 搜索以及根据搜索行为的应用推荐,主要满足用户自然语言的搜索需求.Quixey 是一个完全自动化的移动应用“功能搜索”引擎,它以文本分析、语义分析技术为主,提供移动应用的准确搜索.Quixey 不是简单地根据用户的描述来进行搜索,可以通过 Quixey 定义的函数为用户提供移动应用搜索和发现服务.Quixey 从移动应用商店、论坛、博客、社会化媒体网站和匿名消息来源抓取移动应用的相关信息,并对这些信息进行进一步的抽取、分析、集成,从而提供高质量的功能搜索服务.

上述 3 种方式在一定程度上能够帮助用户快速找到自己所需的移动应用,但还有很大改善和提升的空间.移动应用集成是解决这一问题的有效途径.移动应用集成的主要任务是研究如何把海量的移动应用及其相关信息有效地集成起来,为用户提供高质量的搜索、发现和推荐服务.研究内容主要包括移动应用数据抽取、功能建模、移动应用匹配、移动应用搜索与推荐等.

本文主要对移动应用集成中若干关键研究问题的研究现状进行分析总结,并指出未来的主要研究

方向. 本文第 2 节介绍移动应用集成与传统 Web 数据集成的异同，提出移动应用集成基本框架；第 3 节对移动应用数据抽取相关工作进行分析；第 4 节和第 5 节分别介绍移动应用匹配和移动应用推荐技术；第 6 节指出若干挑战性研究问题；最后对本文内容进行总结.

2　移动应用集成框架

目前关于移动应用集成的研究尚处于起步阶段，其中在移动应用数据抽取方面大都是基于传统的 Web 数据抽取技术，偏重于结构化信息的抽取，对于移动应用功能信息抽取技术的研究还比较少；在移动应用搜索与推荐方面有一些初步研究. 本节首先对 Web 数据集成进行简单介绍，对移动应用集成和 Web 数据集成技术进行对比分析，在此基础上给出移动应用集成的基本框架.

2.1　Web 数据集成

关于 Web 数据集成，大量学者已经作了系统深入的研究，其中刘伟等人[1]对 Deep Web 数据集成进行了综述，提出了 Deep Web 数据集成框架，把集成过程分成了 3 个模块：查询接口生成模块、查询处理模块和查询结果处理模块. 其中查询接口生成模块包括 Web 数据库发现、查询接口模式抽取、Web 数据库分类和查询接口生成 4 个子模块；查询处理模块主要包括 Web 数据库选择、查询转换、查询提交 3 个子模块；查询结果处理模块主要包括结果抽取、结果注释和结果合并 3 个子模块. 文献[2-3]分别对查询接口模式抽取、查询接口的集成进行了研究；文献[4-5]对 Web 数据库的选择、查询转换相关技术进行了分析；文献[6]重点研究了基于视觉的查询结果抽取方法.

查询结果的处理是 Web 数据集成的核心任务. 查询结果处理的主要任务是把来自于多个 Web 数据库的异构的数据以一个统一的形式展示给用户，目前的主要研究工作集中在如何快速准确地从查询结果页面抽取出结构化的查询结果. 目前的 Web 数据抽取主要包括以下几种技术：页面抽取语言、基于 DOM 树的技术、抽取规则推导技术、基于视觉的抽取等.

2.2　移动应用集成与 Web 数据集成的异同

移动应用集成与传统的 Web 数据集成有一些共同点，如属性信息抽取、数据融合等，两者都需要从相应的 Web 页面中抽取出结构化的属性信息，对于不同数据源的数据需要进行消重、融合等. 然而，与 Web 数据集成相比，移动应用集成也有其特殊之处，二者的主要区别见表 2.

表 2　Web 数据集成与移动应用集成的对比

对比项	Web 数据集成	移动应用集成
集成对象	Web 对象本身，如电子商务网站中的产品信息、学术网络中的论文信息、求职网中的职位信息等.	移动应用； 与移动应用相关的动态信息：用户信息、用户评论、社交网络中的分享信息等.
集成目的	把多个数据源的异构信息用统一的结构来表示； 提供一个统一的查询接口.	提供统一的查询接口； 实现移动应用的推荐和匹配； 提供基于功能的查询.
查询方式	基于关键词匹配的查询.	基于关键词匹配的查询； 基于功能的语义查询.
关键问题	查询接口信息集成； 查询转换； 查询结构抽取.	多源信息动态数据抽取； 功能建模； 移动应用匹配.

2.3　移动应用集成基本框架

我们针对移动应用的特点，并结合现有的数据集成技术，提出了移动应用集成框架，如图 1 所示.

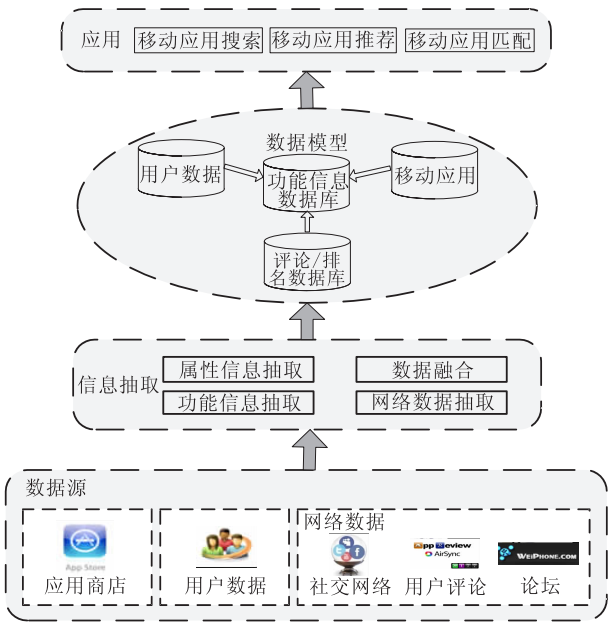


图 1　移动应用集成框架

移动应用集成主要包括 4 个层次：数据源、信息抽取、数据模型和应用. 其中数据源主要包括众多的移动应用商店，以及与移动应用相关的各种网络数据源如用户数据、社交网络、用户评论、论坛信息等. 移动应用商店主要包括移动应用的基本属性信息，该部分信息需要利用相应的数据抽取技术，从多个不同的应用商店中抽取出具有统一格式的结构化信息，并根据实际情况进行数据消重、数据融合等

处理.

移动应用本身的信息是静态信息,而用户数据主要是指用户在使用应用程序的过程中所产生的一系列相关数据,如用户的安装、更新、删除历史,用户使用时间记录,用户在应用程序中的资料信息等.通过这些信息可以分析用户的使用习惯,了解用户的潜在需求,从而为用户提供更好的推荐服务.但是目前该部分信息因隐私问题,不太容易得到.

随着 Web2.0 技术的发展,很多用户都习惯于在网络中分享自己的相关信息,如用户可以在Facebook中与好友分享自己所使用的移动应用程序列表、自己的使用体验、评价等;还有一些针对移动应用的专业论坛,可以供用户之间交流移动应用的使用信息、对应用的评价等,如比较有名的是威锋网.从这些信息中可以全方位了解移动应用,分析移动应用的质量、用户喜爱程度等,对于提高服务质量具有重要作用.

信息抽取主要是从众多数据源中把与移动应用相关的信息抽取出来,主要包括属性信息抽取、数据融合、功能信息抽取和网络数据抽取等.其中属性信息抽取主要是把移动应用相关的结构化信息抽取出来,如应用名称、类别、适合机型、价格等;数据融合主要解决不同数据源中数据的冲突问题;移动应用集成中的属性信息抽取技术和数据融合技术与传统的 Web 数据集成基本相同.功能信息抽取,主要负责从移动应用的描述信息以及与移动应用相关的评论信息中抽取出应用的主要功能,该部分是 Web 数据集成中所没有考虑或者是没有必要考虑的内容.网络数据抽取主要指从与移动应用相关的各种数据源中把所需要的信息抽取出来,如用户评论信息、移动应用的使用排名、用户的评分信息等,该部分主要难点在于相关信息的识别以及数据的动态特性.

模型层主要是把移动应用的基本属性信息、能做什么、做得怎么样、如何使用、用户评价等各种不同的信息以一种合理的方式进行建模,并建立高效的索引,以实现快速和高质量的搜索服务以及其他应用需求.

应用层主要是在已经处理好的移动应用程序数据库的基础上提供相应的服务,如移动应用搜索、移动应用推荐、移动应用匹配等.

3 移动应用数据抽取

移动应用数据抽取是移动应用集成的核心任务之一,同时也是其他任务的基础.在数据抽取方面已经有大量的研究工作,按照不同的标准可以分类不同的类别.按数据来源不同,可以分为基于非结构化数据(文本)的抽取和基于半结构化数据(Web 数据)的抽取;按照自动程度不同,可以分为手动、半自动和全自动的数据抽取.在移动应用集成中,属性信息抽取和功能信息抽取是数据抽取模块的主要目标.属性信息抽取主要是从移动应用所在的 Web 网页中把移动应用的名字、类别、描述等信息抽取出来,功能数据抽取主要是从移动应用的描述信息、论坛信息及用户评论信息中把能够代表移动应用功能的主要短语、句子等抽取出来.

目前在 Web 数据抽取方面已经有了大量的研究工作,其中刘伟、孟小峰等人[1]在《Deep Web 数据集成研究综述》中对 Web 数据抽取技术进行了归纳总结,并按照使用技术的不同进行了分类,主要包括基于 DOM 树的技术[7]、基于模式的技术[8]、页面抽取语言[9]和抽取规则推导技术等.不过文献[1]分析的主要是 2007 年以前的技术,我们不再进行详细介绍,本节主要对近几年提出的一些新的、代表性数据抽取技术进行分析.

D-EEM[10]是一种基于 DOM 树的 Deep Web 实体抽取机制(DOM-tree based entity extraction mechanism for Deep Web).D-EEM 采用基于 DOM 树的自动实体抽取策略,将实体抽取过程分为数据区域定位和实体区域定位两个阶段,从而可以在比较精确的范围内进行实体区域的定位,大大提高了实体抽取的效率;另外,为了提高实体抽取的准确性,在抽取过程中还考虑了 DOM 树内文本内容节点和元素节点的特征.田健伟等人[11]为了能够完整地提取 Deep Web 数据库中的记录,提出了一种基于层次树的数据获取技术.该技术把 Web 数据库建模成一棵层次树,这样 Deep Web 数据的获取问题就可以转化成树的遍历问题.其次通过属性排序和基于属性值相关度的启发规则指导遍历过程提高遍历效率.实验结果表明该方法具有很好的覆盖率和较高的提取效率.OXPath[12]对 XPath 进行了扩展,能够在交互式的网站中支持页面导航和结果数据的抽取.其最大的特点是能够模拟用户的行为,动态获取页面的 CSS 属性信息,并且每次只需处理当前的页面,所以需要的内存空间比较小.

Liu 等人[6]认为传统的 Web 数据抽取技术虽然能够取得较好的抽取效果,但是大多都依赖于 Web 页面编程语言,一旦页面语言发生了改变,抽

取技术也得做相应的改变.为了克服这方面的限制，Liu 等人系统分析了多种结果页面的视觉特征，并使用结果页的视觉特征来进行数据记录和数据项的抽取工作，此方法最大的特点是抽取过程与页面语言种类无关，适合在多语种环境中的使用.

Ferrara 等人[13]从一个新的角度对 Web 数据抽取技术和应用进行了综述.以往的综述论文主要是从数据抽取技术和算法的角度进行分类和描述，而 Ferrara 等人首次从应用的角度对 Web 数据抽取技术进行了分类，深入分析了不同应用领域中 Web 数据抽取技术的相同点和不同点.作者主要从企业应用和社交网络应用两个大的领域进行了分析，并指出了不同应用领域中数据抽取技术存在的挑战性问题.

马安香等人[14]针对重复语义标注和嵌套属性的问题，提出了一种基于结果模式的 Deep Web 数据抽取机制.该机制将数据抽取工作分为结果模式生成和数据抽取两个阶段，在结果模式生成阶段进行属性语义标注，从而解决了重复语义标注问题；在结果模式的基础上提出了一种新的数据抽取方法，很好地解决了嵌套属性问题.

由于移动应用数据往往表达随意，具有不规范性，为了改善移动应用匹配、推荐的效果，需要从这些不规范的、短小的移动应用数据中抽取出其主题或关键词. Zhao 等人[15]主要研究如何从 Twitter 信息中抽取出主题关键短语.Twitter 信息一般都比较短，并且噪音比较多，为了提高抽取质量，作者利用关键词排序、关键短语生成和关键短语排序 3 个阶段来实现.在关键词排序中，基于主题敏感传播算法，对主题 PageRank 算法进行了改进；在关键词排序和关键短语生成的基础上，设计了一个概率短语评分函数，最后利用该评分函数对短语进行排序，取最前面的若干个短语作为关键短语. Yu 等人[16]提出了一种从商品评论中进行主题抽取的方法.作者首先通过预处理，抽取出名词或名词短语，并把这些名词和名词短语作为候选主题；然后计算这些主题的相对词频，如果相对词频低于某个阈值，则过滤掉，不进行后面的处理；最后针对每个候选主题计算其改进的 TF-IDF 值，如果改进的 TF-IDF 值大于某个阈值，则该主题就可以作为最后的结果.另外在进行主题抽取的过程中，为了过滤掉冗余的主题，作者提出了一个主题支持度，如果主题 w_i 的频率小于某个包含 w_i 的短语(w_iw_j)的频率，则 w_i 就可以过滤掉，只把 w_iw_j 作为一个候选主题.

4　移动应用匹配

据我们调研，目前还没有关于移动应用集成相关技术的系统性研究工作，随着移动应用的普及及数量的不断增加，对于移动应用集成的研究具有前瞻性和必要性.移动应用集成中有很多关键性问题需要研究，如信息抽取技术、数据融合、实体识别、自动推荐、应用匹配等.而移动应用匹配在移动应用集成中具有重要意义，是信息集成、推荐和搜索的基础.所以，目前我们主要针对移动应用匹配问题进行研究.

4.1　移动应用属性特点

在移动应用匹配过程中，我们主要是基于移动应用属性来计算其相似度.通过观察我们发现移动应用的名称、描述信息都具有一些特点.

移动应用名称：功能相似的移动应用名称往往包含相同的词，或者包含同义词，有些名称中包含一些复合词如 autolock、shake2mutecall，有些名称中的词不是一个有效的英语单词，仅仅是一个标识如 Okotag、Barcode Scanner.

描述信息的短文本特性：描述信息与传统的文本文档不同，一般都比较短，由若干个句子组成，可以视为短文本.因此，描述信息中单词的共现概率比较低，即使是功能相似的移动应用，可能都不包含共同的词汇或者相同的词比较少.由此得到的文本特征矩阵就比较稀疏，所以传统的向量空间模型无法很好地根据移动应用的描述信息计算其相似度；另外，据我们观察发现，由于移动应用的描述信息一般都是由开发者提供的，所以描述信息的撰写非常不规范，往往包含很多非功能性描述或者说是噪音数据，如广告信息、用户操作指南、平台要求等，这些非功能性描述对于计算移动应用的相似度具有很大的负面影响.因此，为了提高移动应用相似度计算的准确性，我们必须解决稀疏性和噪音问题.

本节后面的内容主要对短文本分析的相关技术和两种移动应用匹配方法进行分析.

4.2　短文本分析

目前已经有很多学者针对短文本进行了大量的研究工作，如短文本的主题发现、短文本的情感分析、短文本相似度计算、分类、聚类等.其中短文本相似度计算和短文本分类技术对移动应用匹配有重要的指导意义，所以本文对最近关于短文本相似度计

算和短文本分类技术方面的研究进行分析总结.

4.2.1 短文本相似度计算

短文本相似度计算的主要任务是用来判断不同的短文本描述之间的相似程度,短文本的相似度越高,说明短文本表达的意思或观点越相似.短文本相似度计算是短文本分析的基础工作,是分类、聚类和主题发现的重要技术之一.

文献[17]主要提出了一种基于概率主题生成模型的短文本相似度计算方法.核心思想是,对于两个待比较的短文本而言,把它们分成两部分,一部分是相同的单词,另一部分是不同的单词;然后在一个给定的短文本集合中,基于 LDA 模型,利用 Gibbs Sampling 方法找出隐含主题及主题的概率分布;接下来在发现的主题分布上计算不同单词的相似度;最后把两者相结合计算总体相似度.该方法能够在一定程度上解决短文本的稀疏性问题,但是其中也存在一些挑战,如隐含主题的个数如何确定,相似度的阈值如何判断等;文献[18]主要针对短文本的稀疏性特点,提出了一种扩充短文本信息的方法.对于每一个短文本,构造一个查询,提交给搜索引擎,然后利用搜索引擎返回的结果来代表短本,这样就可以大大扩充短文本的信息,同时作者提出了一种相似度核函数,用来计算短文本之间的相似度,具有较好的准确性和可扩展性;文献[19]主要是解决句子之间的相似度计算问题,传统的计算方法不具有较好的扩展性,作者提出了一种基于语义网络和统计分析相结合的方法,具有较好的自适应性;文献[20]把短文本的语义信息和统计信息相结合,提出了一种新的短文本模型方法.主要有 3 个步骤:首先基于语义词典如 WordNet 计算出初始的词相似度矩阵;然后以此为基础,对词相似度和短文本相似度进行迭代计算,直至收敛;最后利用得到的词相似度矩阵对原来的文档-词频矩阵进行修正,映射到新的向量空间中,并在新的向量空间中进行短文本相似度的计算,实验表明取得了较好的效果;文献[21]对现有的句子相似度的计算方法进行了分析,包括语法相似度、语义相似度和语用相似度,并提出一种新的基于关键词提取的句子相似度计算方法.通过观察,并不是所有的词对表达句子的意义都起作用,所以作者根据单词的词性、句子语法结构等提取出关键词,并给每个词赋予不同的权重,在此基础上进行相似度的计算;文献[22]从信息检索的角度,对短文本的表示和相似性度量进行了分析,并对各种不同的度量方法进行了对比,包括基于字典的相似度度量、基于词干化和语言模型的相似性度量,并对各种不同的方法进行了实验,分析了各种方法的优势和不足.

4.2.2 短文本分类

由于微博、在线论坛每时每刻都产生大量的数据,这些丰富的数据一方面给人们带来了更大的选择空间,但是面对海量信息,人们如何进行有选择的阅读却遇到了前所未有的巨大挑战.因此对于海量短文本的重新组织分析就显得非常有必要,分类分析是信息挖掘中最重要和最基本的技术之一.

目前短文本的分类算法主要基于有监督学习[23-24].有监督学习必须对训练样本进行手工标注,并且为了确保分类的可扩展性,往往需要标注大量的样本作为训练集.然而大量样本的标注费时费力,特别是在短文本当中,由于其海量性、不规范性,短文本中的标注问题更为突出.

文献[24]主要针对短文本的稀疏性和描述信号弱的特点,提出了一种基于特征扩展的中文短文本分类方法.该方法主要利用关联规则挖掘算法挖掘训练集特征项和测试集特征项之间的共现关系,然后利用得到的关联规则对测试文档集中的词语进行特征扩展,在此基础上进行短文本分类;文献[25]针对短文本的稀疏性特点,提出了另外一种新的解决方法,针对每一个特定领域的分类问题,首先选择一个足够大规模的外部数据源,并从中发现其中的隐含主题,最后利用这些隐含主题和小规模的标注训练集进行分类;文献[26]中指出独立主成分分析(ICA)在很多情况下能够改善文本分类的效果,但是由于短文本的稀疏性,它们之间相同的词很少,所以直接在短文本上进行独立主成分分析效果不佳.基于此,作者利用潜在语义分析(LSA)对短文本进行数据预处理,然后在此基础上再利用主成分分析,取得了不错的效果;文献[27]主要解决的是 Twitter 消息的分类问题,作者通过一定的算法,把每个 Twitter 消息映射到最相似的 Wikipedia 页面上,然后利用此页面来代表 Twitter 消息,并进行分类,实验表明该方法比单纯的基于字符串编辑距离或 LSA 的效果好;以往的分类研究中每一个短文本只赋予一个类别,而实际上,一个文本有可能包含多个不同的主题,文献[28]主要研究了短文本的多值分类问题;为了能够对海量 Twitter 消息进行重新组织,便于用户选择和浏览,文献[29]针对 Twitter 消息的特点提出了一个新的分类方案.作者首先通过观察和分析,利用贪婪算法选择了 8 个特征,并将这 8 个特征和传统的词袋子方法进行了对比实验,结果表明作者提出的方法具有较高的准确性.

4.3 基于 WordNet 的移动应用匹配

该方法主要是基于移动应用的描述信息计算相似度，把每一个 App 看成是一个由描述信息表示的文档，利用传统的向量空间模型(VSM)[30]进行计算. 为了解决文档-词频矩阵的稀疏性问题，可利用语义词典 WordNet 来扩充 App 的描述信息. 具体实现过程如下：

$a_1,a_2,\cdots,a_M$分别表示 M 个 App 的描述信息，描述信息经过分词、去除停用词和词干化等处理以后，共得到由 N 个不同的词组成的集合 $T=\{t_1,t_2,\cdots,t_N\}$，$|T|=N$；最后得到文档词频矩阵 $\boldsymbol{W}$.

$$\boldsymbol{W}=\begin{pmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \ddots & \vdots \\ w_{M1} & \cdots & w_{MN} \end{pmatrix} \tag{1}$$

其中，每一行代表一个 App，每一列代表一个单词，每一元素 w_{ij} 表示第 j 个单词在第 i 个 App 描述中的权重，计算方法如下：

$$w_{ij}=(1+\log tf_{i,j})\times \log_{10} N/\mathrm{d}f_j \tag{2}$$

然后基于 WordNet，计算词与词之间的语义相似度，得到词语的相似度矩阵 $\boldsymbol{Q}$.

$$\boldsymbol{Q}=\begin{pmatrix} q_{11} & \cdots & q_{1N} \\ \vdots & \ddots & \vdots \\ q_{N1} & \cdots & q_{NN} \end{pmatrix} \tag{3}$$

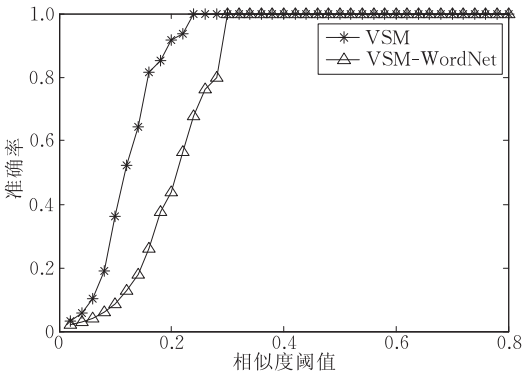
$$\hat{\boldsymbol{W}}=\boldsymbol{W}\times\boldsymbol{Q}=\begin{pmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & \ddots & \vdots \\ w_{M1} & \cdots & w_{MN} \end{pmatrix}\begin{pmatrix} q_{11} & \cdots & q_{1N} \\ \vdots & \ddots & \vdots \\ q_{N1} & \cdots & q_{NN} \end{pmatrix}=\begin{pmatrix} \hat{w}_{11} & \cdots & \hat{w}_{1N} \\ \vdots & \ddots & \vdots \\ \hat{w}_{M1} & \cdots & \hat{w}_{MN} \end{pmatrix} \tag{4}$$

通过上述运算，文档-词频矩阵的非零元素增多，稀疏度降低. App 之间的相似度在转换后的向量空间中利用式(5)进行计算.

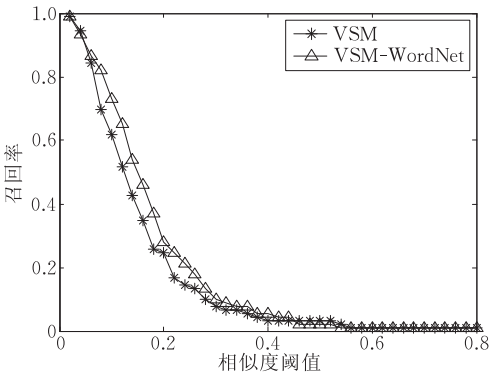
$$Sim(a_1,a_2)=\frac{\langle a_1\circ a_2\rangle}{\|a_1\|\times\|a_2\|}=\frac{\sum_{i=1}^{N}\hat{w}_{a_1 i}\times\hat{w}_{a_2 i}}{\sqrt{\sum_{i=1}^{N}\hat{w}_{a_1 i}^2}\times\sqrt{\sum_{i=1}^{N}\hat{w}_{a_2 i}^2}} \tag{5}$$

我们人工构建了一个小规模的测试数据集，对 100 个 App 进行了人工判断，发现其中共有 89 对相似的 App，对此分别利用 VSM 模型和基于 WordNet 的 VSM 模型进行计算. 实验结果用准确率、召回率、F-1 进行衡量.

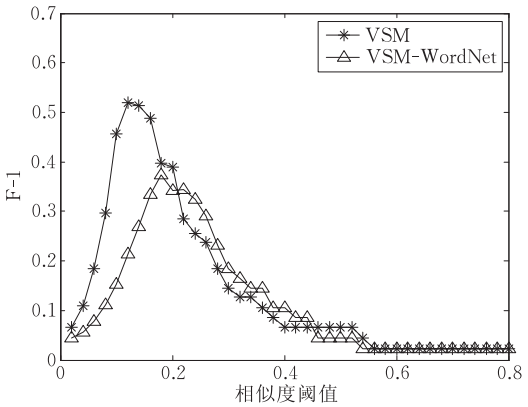
从图 2 可以看出，利用语义词典 WordNet 可以增加 App 之间的相似度，从而提高了召回率，但是准确率却大大下降. 通过分析，准确率下降的主要原



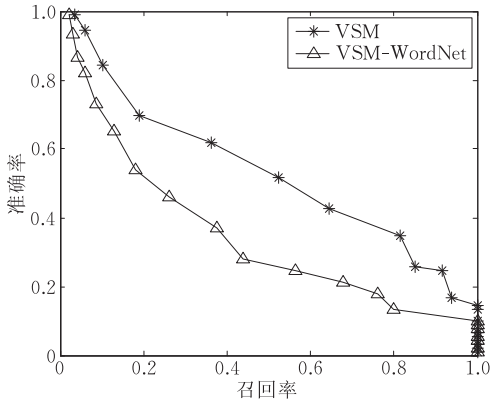
(a) 不同相似度阈值下的准确率



(b) 不同相似度阈值下的召回率



(c) 不同相似度阈值下的F-1值



(d) 准确率与召回率的关系

图 2　基于 WordNet 的移动应用匹配

因是由于 App 描述信息中存在噪音数据.因此单纯利用语义词典无法很好地解决 App 的相似度计算问题,必须想办法消除 App 描述中的噪音信息.

4.4 基于特征词提取的移动应用匹配

为了改善移动应用匹配的效果,需要识别出 App 描述信息中的特征词,这些特征词能够体现 App 的功能,从而把描述信息中的非功能性信息或者噪音数据过滤掉.通过深入观察分析,我们选择 5 个特征作为判断一个词是否是特征词的依据,分别是 termPOS、locInDes、isNameTerm、locRelative-ToName、termFreq,具体说明如表 3 所示.

表 3 特征词列表

序号	特征	说明
1	termPOS	单词词性
2	locInDes	单词在 App 描述中的位置
3	locRelativeToName	单词与 App Name 的相对位置
4	isNameTerm	是否是 App Name 中的单词
5	termFreq	单词出现的频率

我们把特征词的判断问题看成是一个分类问题,主要通过以下几个步骤实现:(1)针对 100 个 App 的描述信息进行手工标注,一共标注了 2625 个单词,如果某个单词在一个 App 中是特征词,则标注为 1,否则标注为 0;(2)计算出每个单词的所有的特征值;(3)以这些标注数据作为训练集,得到一个分类模型;(4)利用该分类模型去判断其他的词是否是特征词.

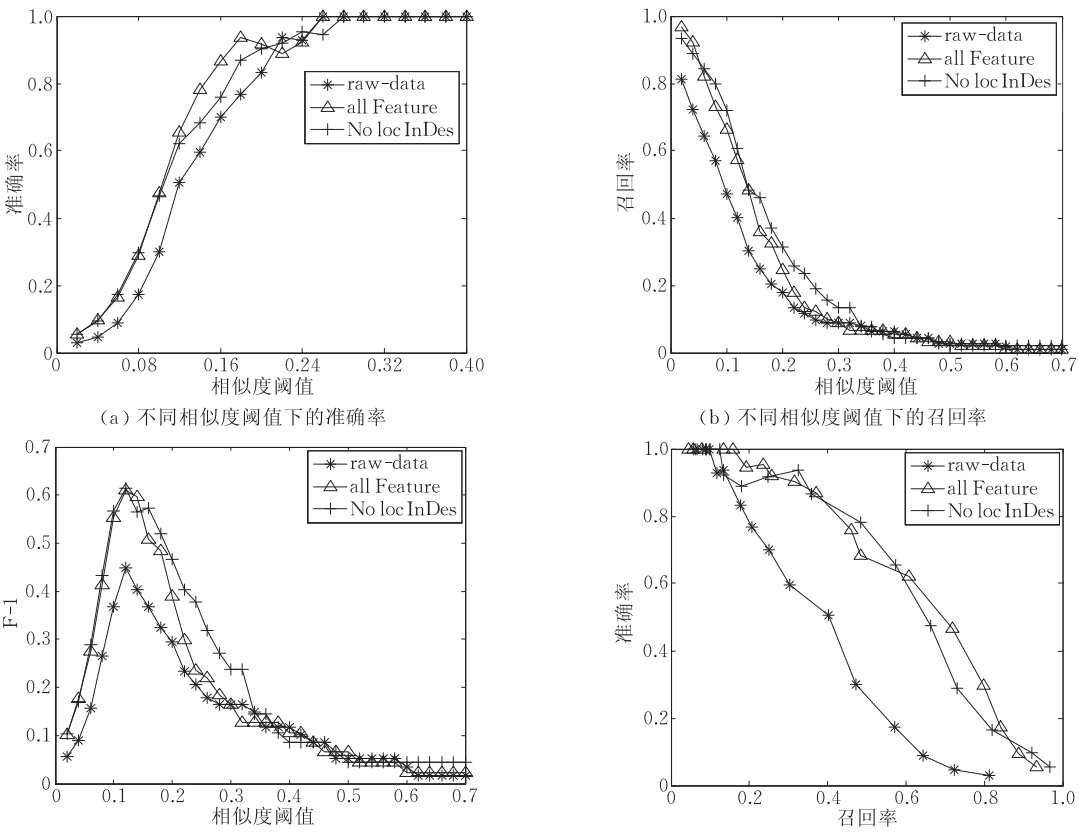
特征词分类实验设置:在 2625 个标注数据中选择 2525 个作为训练集,100 个作为测试集,分别采用朴素贝叶斯(Naïve Bayesian)和支持向量机(SVM)方法进行实验,分类结果如表 4 所示.

表 4 分类正确率

序号	算法	特征	正确率/%
1	Naïve Bayesian	all	54
2	SVM	all	66
3	SVM	No termPOS	58
4	SVM	No locInDes	73
5	SVM	No locRelativeToName	68
6	SVM	No isNameTerm	64
7	SVM	No termFreq	66

从表 4 我们可以看出,朴素贝叶斯分类的正确率比较低,另外去除 locInDes 之后利用 SVM 分类,正确率最高,也就是说 locInDes 对于特征词的判断具有一定的负面作用,但对于最终的 App 相似度计算结果的影响还不确定,所以我们采用 SVM 方法分别在所有特征和去除 locInDes 以后的子集上进行了实验.最后以所有的特征词为向量空间来计算 App 的相似度,实验结果表明,取得了较好的效果.

从图 3 可以看出,经过特征词提取以后,准确率



(a) 不同相似度阈值下的准确率

(b) 不同相似度阈值下的召回率

(c) 不同相似度阈值下的F-1值

(d) 准确率与召回率的关系

图 3 基于特征词提取的移动应用匹配

和召回率均有所提高，并且在不考虑 locInDes 的情况下效果更好，说明单词在描述中的位置对单词是否是特征词没有太大贡献，并且对相似度计算具有负面影响.

上述两种方法都是基于 App 的描述信息进行计算的，以后将把 App 的名称、类别及其他相关信息也考虑进去，效果可能会更好.

5　移动应用推荐与搜索

随着移动应用数量的不断增加，如何帮助用户快速找到想要的应用成了一个亟待解决的问题，部分学者对移动应用的推荐技术进行了研究. Shi 等人[31]首先分析了传统推荐模型存在的不足之处，如以记忆为基础(Memory-based Models)的协同过滤模型(包括以用户为基础的协同过滤和以项目为基础的协同过滤)对经常出现或比较流行的项目推荐效果比较好，但是对于使用不是很频繁的项目推荐效果比较差；隐语义模型(Latent Factor Models)的推荐准确率比较低. 针对上述两种推荐模型存在的不足之处，作者提出了一种新的推荐模型——基于主成分分析的模型(PCA-based model). 该模型首先利用主成分分析技术从数据中找到主要的特征，然后在主要特征的基础上再利用协同过滤模型进行推荐. 其主要优点是对于不是很流行的移动应用具有较好的推荐准确率. Woerndl 等人[32]针对移动应用提出了一种基于情景感知的混合推荐系统. 该推荐系统以传统的协同过滤技术为基础，把情景因素考虑进来，从用户、项目和情景 3 个维度进行计算，大大提高了推荐准确率. 但是目前考虑的情景还比较少，主要是依据其他用户在某个位置的移动应用安装和使用情况进行推荐，以后将考虑更多的情景因素. Karatzoglou 等人[33]结合情景信息，也提出了一个新的移动应用推荐模型 Djinn 模型，该模型主要考虑是把隐式反馈数据考虑进来，利用张量分解技术对 Djinn 模型进行优化，实验结果表明 Djinn 模型的平均准确率(MAP)要比不考虑情景信息的模型高出 28%. Yin 等人[34]认为移动应用的推荐和其他领域的推荐有一个不同之处在于：除了推荐用户感兴趣的移动应用外，还需要针对用户已经有的移动应用推荐可以替代的、新的移动应用. Yin 等人[34]认为已有的移动应用拥有一个实际满意度值 AV(Actual Satisfactory Value)，新的移动应用拥有一个吸引度值 TV(Tempting Value)，用户是否更换旧的应用，取决于 AV 和 TV 的大小. 作者以用户的使用日志为基础数据，把 AV 和 TV 作为两个隐含参数，提出了一个 AT 模型，计算出每个应用的 AV 和 TV 值，并设计了 AT 排序函数. 实验表明，AT 模型的推荐效果远好于传统的协同过滤技术和以内容为基础的过滤技术，如果能将 AT 模型和其他模型相结合，效果会更好. Yan 等人[35]认为以往的移动应用推荐系统大都利用用户的下载历史和用户评价，实际上用户下载了一个应用，并不能真正代表用户喜欢该应用，而用户的评价往往又比较稀疏，推荐效果不佳. 因此他们把用户的使用日志数据和基于项目的协同过滤技术相结合，提出了一种个性化的移动应用推荐技术 AppJoy. Zhu 等人[36]对移动应用的分类问题进行了研究. 为了提高分类的准确性，作者对移动应用的特征信息进行了扩展：一是利用搜索引擎来扩展文本特征；二是从用户的使用记录中提取情景特征，最后把这些特征综合起来，利用最大熵模型训练出了一个移动应用分类器. 实验结果表明其分类准确率要高于基于词向量的应用分类器(Word Vector based App Classifier)和基于隐含主题的应用分类器.

随着移动应用数量的不增加，移动应用搜索将越来越重要. 移动应用搜索与传统的 Web 搜索有相似之处，但也有特殊之处. 移动应用搜索对搜索结果的质量要求更高，需要返回最能够满足用户需求的少数应用，而不需要返回大量的结果；另外在移动应用搜索中，传统的以关键词为基础的搜索技术无法满足新的查询需求，因为用户往往不能够准确给出应用的名称，只能大概给出应用的功能、特点，在这种情况下，如何能够准确分析出用户的查询意图并提供满意的结果将变得非常具有挑战性；移动应用搜索结果的排名也有特殊之处，除了考虑搜索结果与用户查询之间的相关性之外，还需要考虑应用的质量、受欢迎程度等其他因素. 因此，功能搜索或者是语义搜索将是解决移动应用搜索的一个有效途径. 但是目前还没有比较好的解决方案.

6　移动应用集成面临的挑战

目前，关于移动应用集成技术的研究还处于刚刚起步阶段，并且由于移动应用本身的特点，在移动应用集成中存在一系列挑战，主要包括多源信息集成、功能信息抽取和建模、移动应用匹配和移动应用排名等.

6.1 多源信息集成

移动应用集成的数据对象除了移动应用的基本属性之外,还包括与移动应用相关的其他动态信息:用户信息、用户评论、社交网络中的分享信息等.这些信息对改善移动应用的搜索和推荐效果具有重要作用.然而这些信息往往存在于不同的数据源中,如移动应用的基本属性信息大都存在于各大应用商店或者部分移动应用集成网站,而相关的用户评论、社交网络分享信息等则存在于其他网站中,不同的数据源具有不同的页面结构,如何设计具有自适应能力的抽取方法是一个巨大的挑战.其次移动应用相关的数据源大都具有 Web2.0 的特征,所以数据源中页面的结构经常会发生变化,如何使得数据抽取方法在页面结构发生变化时仍能够继续工作也是一个重要的研究内容.关于多源信息的集成,部分学者已经做了研究.Spiegel 等人[37]和 Szomszor 等人[38]为了改善电影推荐效果,尝试将 IMDB 和 Netflix 的数据进行集成.IMDB 是一个在线的电影信息共享网站,它允许用户对影片添加标签,来描述影片的演员信息、情节、故事地点等.NetFlix 是一个在线视频租赁网站,用户可以对看过的视频打分.Spiegel 等人[37]和 Szomszor 等人[38]将 IMDB 的标签信息和 Netflix 的打分信息进行集成,大大提高了推荐的效果.

6.2 功能信息抽取与建模

功能信息抽取也是一个极具挑战性的问题,对移动应用的搜索效果具有重要影响.传统的 Web 数据抽取技术可以从半结构化数据中抽取出与应用相关的属性信息,如名称、类别、描述、价格等;但是移动应用的功能性信息更为重要,比如应用能实现哪些功能?做得怎么样?如何使用等?这些功能性信息是功能搜索的基础,对提高功能搜索的质量至关重要.然而,功能性信息往往隐藏在移动应用的描述信息、用户评论等非结构化信息中,传统的 Web 数据抽取技术无法从非结构化信息中抽取相应的结构化信息.虽然已经有一些自然语言处理的相关技术可以从非结构化信息中进行信息提取,但是还不能直接应用于此,主要原因在于移动应用的描述信息以及相关用户评论等具有自己的特点,如文本短小、语法不规则等.

移动应用集成的主要目的之一就是提供高质量的搜索服务,使用户能够得到真正满足实际需求的结果.移动应用搜索和传统搜索的最大区别在于:传统搜索主要是以关键词匹配为主,而关键词匹配在移动应用搜索中效果非常不好,目前几大移动应用商店提供的搜索功能都不能令人满意.目前已有很多公司涉足 App 搜索市场,如提供功能搜索的 App 搜索引擎 Quixey,百度也推出了 App 搜索平台.但是目前各公司所采用的 App 搜索技术并没有对外公布,学术界关于 App 搜索还没有相关的研究.人们在搜索应用时往往不知道其准确名字,希望搜索出能够完成某种任务、具备某种功能的软件,如观看 NBA 比赛、视频编辑、寻找最近的超市等,针对这些查询,传统搜索无法提供很好的结果.功能建模是解决这一问题的核心.

功能建模的主要目的是提供高质量的搜索服务,能够实现基于功能的搜索.在数据抽取阶段,通过各种抽取技术,得到了移动应用的基本属性信息、功能信息、评论信息以及用户数据,功能建模主要是以功能为核心,设计一种合适的数据模型,把上述各种信息进行有效的表示、组织与存储,数据空间技术和语义网技术是功能建模可以借鉴和参考的两个技术;同时,为了提高搜索的效率,必须根据新的数据模型的特点设计高效的索引策略.

6.3 移动应用匹配

移动应用匹配主要是用来判断两个应用程序在功能上是否相似,是实现移动应用迁移、移动应用推荐的基础,是一个重要的研究内容,有很多的应用场景.

移动应用匹配与实体识别具有一定的相似性.实体识别主要用来判断两个不同的数据记录是否代表同一个实体,目前已经有大量的相关研究工作.按照所使用的技术不同可以分为以下几类:概率匹配模型[39]、监督和半监督学习方法[40]、主动学习技术[41]、基于距离的技术、基于规则的方法和无监督学习的方法.实体识别主要是基于实体的属性信息进行相似度比较,而移动应用匹配过程中,除了考虑属性信息的相似度之外,应用程序的功能相似度更为重要,所以传统的实体识别技术并不能直接应用于移动应用匹配.

首先,属性选择是移动应用匹配的首要任务.每个应用都有很多属性信息,如名称、类别、机型、价格、功能描述等,然而并不是所有的属性都对应用匹配起正面作用,所以需要从众多的属性中选出能反映应用功能相似性的属性;

其次,短文本的相似度计算也是一个极具挑战性的研究内容.目前已经有一些研究者对网络短文本进行了一些研究,包括基于语义的方法、基于概率主题模型的方法[17]、基于特征扩展的方法等[24].但

是这些方法并没有考虑移动应用描述信息的特定表达方式，所以无法取得较好的计算效果.

另外，在进行移动应用匹配的过程中，除了考虑应用本身的功能相似性之外，往往还需要考虑用户的使用习惯、个人爱好等信息；同时还需要考虑应用与用户已有的应用之间的相互协作关系，应用彼此之间的相互影响等. 从而为用户提供更加智能和完善的服务.

6.4　移动应用排名

在移动应用集成系统中，最终的目的是为用户提供移动应用的搜索和推荐服务，因此移动应用的排名也是一个重要的研究问题. 应用的排名除了考虑与查询关键词的匹配程度之外，还需要考虑其他相关信息，如用户的偏好、用户查询意图等，需要将这些信息综合考虑，设计一个合理有效的排名函数. 同时由于网络信息具有时变性，现在被用户喜爱的应用，随着时间的推移可能变得不那么受人喜爱，应用的排名可能也会随时间发生变化，所以如何对这些信息进行动态的更新维护，也是一个颇具挑战性的问题.

6.5　移动应用内数据集成与搜索

目前本文中所关注的集成对象主要是移动应用的属性信息以及其他相关信息，如用户评论、社交网络分享信息等，这些可以认为是移动应用的外在信息. 然而，对于用户来讲，移动应用内部所包含的内容更丰富、价值更大. 如果能够把众多移动应用内部的信息有效地集成起来，为用户提供统一的搜索服务，对用户将具有重要的意义. 与传统的网页数据相比，移动应用内部信息的集成与搜索具有一些新的挑战. 信息获取比较困难：移动应用内的信息往往被包上了外壳，无法使用传统的搜索爬虫技术直接抓取；数据格式的异构性：不同的移动应用，其内部的数据格式往往不一样，并且存在大量的噪音数据，其数据抽取方式与网页数据抽取相比更为复杂.

7　结束语

目前移动互联网的流量快速增加，未来必将超过传统互联网，而移动应用逐渐成为移动互联网的主要接入方式. 为了争夺用户，电信运营商、手机制造商、互联网服务提供商以及各个不同的企业纷纷推出自己的移动应用，移动应用数量呈现爆炸式增长. 然而随着移动应用数量的不断增加，给移动应用的搜索和推荐带来了很大的困难. 移动应用集成是改善移动应用搜索和推荐效果的一个有效途径. 目前关于移动应用集成，学术界还没有开展系统深入的研究. 本文提出了移动应用集成的基本框架，对其中的关键技术如数据抽取、移动应用匹配、移动应用推荐等进行了分析，对现有的工作进行了归纳总结；最后指出了移动应用集成中的若干挑战性问题. 未来移动应用的数量将持续增加，成为人们获取信息的主要途径，然而其数量的增加也必将带来一系列挑战，有很多问题值得研究. 我们对移动应用的集成、匹配、推荐等技术进行了分析，希望能为相关研究人员提供参考.

参　考　文　献

[1] Liu Wei, Meng Xiao-Feng, Meng Wei-Yi. A survey of Deep Web data integration. Chinese Journal of Computers, 2007, 30(9): 1475-1489(in Chinese)
(刘伟，孟小峰，孟卫一. Deep Web 数据集成研究综述. 计算机学报，2007，30(9)：1475-1489)

[2] Zhang Z, He B, Chang K C. Understanding Web query interfaces: Best-effort parsing with hidden syntax//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'04). Paris, France, 2004: 107-118

[3] Wu W, Yu C, Doan A, Meng W. An interactive clustering-based approach to integrating source query interfaces on the Deep Web//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'04). Paris, France, 2004: 95-106

[4] Yu C, Philip G, Meng W. Distributed top-*n* query processing with possibly uncooperative local systems//Proceedings of the 29th International Conference on Very Large Data Bases (VLDB'03). Berlin, Germany, 2003: 117-128

[5] Zhang Z, He B, Chang K C. Light-weight domain-based form assistant: Querying web databases on the fly//Proceedings of the 31st International Conference on Very Large Data Bases(VLDB'05). Trondheim, Norway, 2005: 97-108

[6] Liu W, Meng X, Meng W. ViDE: A vision-based approach for Deep Web data extraction. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(3): 447-460

[7] Liu L, Pu C, Han W. XWRAP: An XML-enabled wrapper construction system for web information sources//Proceedings of the 16th International Conference on Data Engineering (ICDE'00). San Diego, USA, 2000: 611-621

[8] Meng X, Lu H, Gu M. SG-WRAP: A schema-guided wrapper generator//Proceedings of the 18th International Conference on Data Engineering(ICDE'02). San Jose, USA, 2002: 331-332

[9] Arocena G, Mendelzon A. WebOQL: Restructuring documents, databases, and webs//Proceedings of the 14th International Conference on Data Engineering (ICDE'98). Orlando, USA, 1998: 24-33

[10] Kou Yue, Li Dong, Shen De-Rong, et al. D-EEM: A DOM-tree based entity extraction mechanism for Deep Web. Journal of Computer Research and Development, 2010, 47(5): 858-865(in Chinese)
(寇月, 李冬, 申德荣等. D-EEM: 一种基于 DOM 树的 Deep Web 实体抽取机制. 计算机研究与发展, 2010, 47(5): 858-865)

[11] Tian Jian-Wei, Li Shi-Jun. Retrieving Deep Web data based on hierarchy tree model. Journal of Computer Research and Development, 2011, 48(1): 94-102(in Chinese)
(田建伟, 李石君. 基于层次树模型的 Deep Web 数据提取方法. 计算机研究与发展, 2011, 48(1): 94-102)

[12] Furche T, Gottlob G, Grasso G, et al. OXPath: A language for scalable data extraction, automation, and crawling on the Deep Web. VLDB Journal, 2013(2): 47-72

[13] Ferrara E, Fiumara G, Baumgartner R. Web data extraction, applications and techniques: A survey. ACM Transaction on Computational Logic, 2010, V(N): 1-20

[14] Ma An-Xiang, Zhang Bin, Gao Ke-Ning, et al. Deep Web data extraction based on result pattern. Journal of Computer Research and Development, 2009, 46(2): 280-288 (in Chinese)
(马安香, 张斌, 高克宁等. 基于结果模式的 Deep Web 数据抽取. 计算机研究与发展, 2009, 46(2): 280-288)

[15] Zhao W, Jiang J, He J, et al. Topical keyphrase extraction from Twitter//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11). Oregon, Portland, 2011: 379-388

[16] Yu L, Duan X, Tian S, Guo H. Topic extraction based on product reviews. Journal of Computational Information Systems, 2013, 9(2): 773-780

[17] Quan X, Liu G, Lu Z, et al. Short text similarity based on probabilistic topics. Knowledge and Information Systems, 2010, 25(3): 473-491

[18] Sahami M, Heilman T. A Web-based kernel function for measuring the similarity of short text snippets//Proceedings of the 15th International Conference on World Wide Web (WWW'06). Edinburgh, UK, 2006: 377-386

[19] Li Y, McLean D, Bandar Z, O'Shea J, Crockett K. Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(8): 1138-1150

[20] Liu W, Quan X, Feng M. A short text modeling method combining semantic and statistic information. Information Sciences, 2010, 180(20): 4031-4041

[21] Pei Jing, Bao Hong. Application of Chinese sentence similarity computation in FAQ. Computer Engineering, 2009, 35(17): 46-48(in Chinese)
(裴婧, 包宏. 汉语句子相似度计算在 FAQ 中的应用. 计算机工程, 2009, 35(17): 46-48)

[22] Metzler D, Dumais S, Meek C. Similarity measures for short segments of text//Proceedings of the 29th European Conference on Information Retrieval (ECIR'07). Rome, Italy, 2007: 16-27

[23] Ning Ya-Hui, Fan Xing-Hua, Wu Yu. Short text classification based on domain word ontology. Computer Science, 2009, 36(3): 142-145(in Chinese)
(宁亚辉, 樊兴华, 吴渝. 基于领域词语本体的短文本分类. 计算机科学, 2009, 36(3): 142-145)

[24] Wang Xi-Wei, Fan Xing-Hua, Zhao Jun. Method for Chinese short text classification based on feature extension. Journal of Computer Applications, 2009, 29(3): 843-845(in Chinese)
(王细薇, 樊兴华, 赵军. 一种基于特征扩展的中文短文本分类方法. 计算机应用, 2009, 29(3): 843-845)

[25] Phan X, Nguyen L, Horiguchi S. Learning to classify short and sparse text & Web with hidden topics from large-scale data collections//Proceedings of the 17th international conference on World Wide Web(WWW'08). Beijing, China, 2010: 91-100

[26] Pu Q, Yang G. Short-text classification based on ICA and LSA//Proceedings of the 3rd International Conference on Advances in Neural Networks(ISNN'06). Chengdu, China, 2006: 265-270

[27] Genc Yegin, Sakamoto Yasuaki, Nickerson Jeffrey V. Discovering context: Classifying tweets through a semantic transform based on Wikipedia//Schmorrow D D, Fidaopiastis C M. Foundations of Augmented Cognition. Directing the Future of Adaptive Systems. Lecture Notes in Computer Science 6780. Springer, 2011: 484-492

[28] Heß A, Dopichaj P, Maaß C. Multi-value classification of very short texts//Proceedings of the 31st Annual German Conference on Advances in Artificial Intelligence (KI'08). Kaiserslautern, Germany, 2008: 70-77

[29] Sriram B, Fuhry D, Demir E, et al. Short text classification in twitter to improve information filtering//Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'10). Geneva, Switzerland, 2010: 841-842

[30] Guo Q. The similarity computing of documents based on VSM//Proceedings of the 2nd International Conference on Network-Based Information Systems (NBiS'08). Turin, Italy, 2008: 142-148

[31] Shi K, Ali K. GetJar mobile application recommendations with very sparse datasets//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'12). Beijing, China, 2012: 204-212

[32] Woerndl W, Schueller C, Wojtech R. A hybrid recommender system for context-aware recommendations of mobile applications//Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop (ICDEW'07). Istanbul, Turkey, 2007: 871-878

[33] Karatzoglou A, Baltrunas L, Church K, Böhmer M. Climbing the App wall: Enabling mobile App discovery through context-aware recommendations//Proceedings of the 21st ACM International Conference on Information and Knowledge Management(CIKM'12). Maui, USA, 2012: 2527-2530

[34] Yin P, Luo P, Lee W, Wang M. App recommendation: A contest between satisfaction and temptation//Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13). Rome, Italy, 2013: 395-404

[35] Yan B, Chen G. AppJoy: Personalized mobile application discovery//Proceedings of the 9th International Conference on Mobile Systems (MobiSys'11). Washington, USA, 2011: 113-126

[36] Zhu H, Cao H, Chen E, et al. Exploiting enriched contextual information for mobile app classification//Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12). Maui, USA, 2012: 1617-1621.

[37] Spiegel S, Kunegis J, Li F. Hydra: A hybrid recommender system [cross-linked rating and content information]//Proceedings of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management (CNIKM'09). Hong Kong, China, 2009: 75-80

[38] Szomszor M, Cattuto C, Alani H, et al. Folksonomies, the semantic web, and movie recommendation//Proceedings of the 4th European Semantic Web Conference, Bridging the Gap Between Semantic Web and Web 2.0 (ESWC'07). Innsbruck, Austria, 2007

[39] Newcombe H, Kennedy J, Axford S, James A. Automatic linkage of vital records. Science, 1959, 130(3381): 954-959

[40] Cochinwala M, Kurien V, Lalk G, Shasha D. Efficient data reconciliation. Information Sciences, 2001, 137(1-4): 1-15

[41] Cohn D, Atlas L, Ladner R. Improving generalization with active learning. Machine Learning, 1994, 15(2): 201-221

MA You-Zhong, born in 1981, Ph. D. candidate. His research interests include Web data and cloud data management.

MENG Xiao-Feng, born in 1964, professor, Ph. D. supervisor. His research interests include Web data management, cloud data management, mobile data management, XML data management, flash-aware DBMS and privacy preserving.

JIANG Da-Xin, born in 1975, Ph. D., researcher of Microsoft Research Asia. His research interests include data mining and information retrieval.

Background

Mobile Application store paradigm achieved big success since it was firstly proposed by apple in 2008, motivated by the successful case many companies such as handset manufacturers, carriers and Internet Service Providers also lunched their own Application store. So Mobile Application store has become a new model of internet development. While with the explosion of the mobile applications, a series of problems come up related with mobile application search and discovery, it is becoming a boring thing for the users to find their desired applications. In order to resolve these problems, it is essential to do some research on mobile application integration. Mobile applications have many unique features compared with the traditional web pages, so several challenging research problems exist in the mobile application integration, such as mobile application data extraction, mobile application functional search and mobile application matching.

In this paper we analyzed the characteristics of mobile applications and make some summarization, proposed a mobile application integration framework aiming to resolve the problems arising from the explosion of mobile applications, made a deep insight analysis about some key issues. We want to help much more researchers pay attentions to the mobile application integration related issues that need to be addressed. We have done some preliminary works and mainly focus on the mobile application similarity computation. Mobile application similarity can display how similar two applications are in functionality, and it is the core issue to deal with the application search and application match. We do some experiments and the results show that our method works well.

This work is partially supported by the grants from the National Natural Science Foundation of China (Nos. 61070055, 91024032, 91124001); the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University (No. 11XNL010); National High Technology Research and Development Program (863 Program) of China (Nos. 2012AA010701, 2012AA011001, 2013AA013204).