热门城市旅游酒店的个性化推荐系统

计算机科学技术学院 2016级云分析班 陈爽 20161104552

指导教师 苗忠义 讲师 穆文龙 讲师

摘要 随着信息技术和互联网的飞速发展，它给人们带来了越来越多的方便，同时也在改变着人们的生活方式。很多人习惯选择在网上购买很多东西，例如：在网上预订酒店、机票和电影票等。然而人们在外出旅行的时候对目的地是陌生的，从而对酒店的需求也是各有不同。所以开发出一套酒店的个性化推荐系统具有很重要的意义。

首先使用爬虫在对某酒店网站进行酒店数据爬取，爬取到不同热门旅游城市的数据建立不同的城市数据表。再经过Hadoop分布式环境进行实现对数据的分布式处理。推荐系统中应用环境功能分为两个：①默认为用户推荐当地最热门酒店没有筛选条件；②根据用户对酒店的不同需求，为用户呈现最终筛选酒店。

关键词 推荐系统；爬虫；Hadoop；数据库

**Personalized Recommendation System of Popular City Tourist Hotels**

Computer College 2016 cloud analysis class Chen Shuang 20161104552

Directed by Miao Zhong Yi Lecturer Mu Wen long Lecturer

**Abstract** With the rapid development of information technology and Internet, it brings more and more convenience to people, but also changes people's way of life. Many people are used to buying many things online, such as booking hotels, air tickets and movie tickets. However, when people travel, they are unfamiliar with the destination, so they have different needs for the hotel. Therefore, it is of great significance to develop a personalized recommendation system for hotels.

First of all, we use the crawler to crawl the hotel data from a hotel website, and build different city data tables by crawling the data from different popular tourist cities. Then through the Hadoop distributed environment to realize the distributed processing of data. The application environment function in the recommendation system is divided into two parts: ① the default is that the user recommends the most popular local hotel without screening conditions; ② according to the different needs of the user for the hotel, the user presents the final screening hotel.

**Keywords** Recommender System；Reptile；Hadoop；Data Base

# 第一章 绪论

# 推荐系统简介

随着信息技术和互联网的飞速发展，让人们无论何时何地都能够获得海量的信息，让人们从信息匮乏的时代走向了信息爆炸的时代。面对大量的数据，人们无法从中获取到对其有用或者感兴趣的信息。因此一个好的过滤系统给人们的生活带来的帮助是很大的。

推荐系统是根据用户的要求和兴趣以及喜好等为用户推荐他可能会感兴趣的个性化服务系统。推荐系统可以根据用户的需求，明确的为用户推荐出他较为感兴趣的信息内容，极大的方便了用户所需信息的获取。

# 课题背景和意义

随着互联网技术的兴起，极大地丰富了人们各个方面的生活，同时也在默默的改变人们的生活方式。通过相关调查数据显示，人们外出旅游度假时，选择互联网去了解目的地的信息已经成为潮流，人们在各大酒店平台上去了解酒店的详细信息，例如：地理位置、酒店价格和舒适程度等。有70%的用户群体在外出旅游的行程规划上，对酒店的选择还是以网络信息为主要依据。通过发达的互联网技术，人们很轻松的能够找到目的地的各种信息，能够了解到旅游的相关服务和产品，可以进行在线预订，拉近了用户与产品之间的距离。在2014年CNNIC发布的《中国在线旅游预订市场研究报告》中的数据显示，截止到2015年初，已有2.2亿人选择使用互联网来预订酒店、飞机票、火车票，以及其他的生活用品，相比2013年增长了4096万人，同比增长率高达22.6%，用户使用率也增长到34.3%，然而在2012年，这一数据仅为7.9%。如此惊人的增长量可以看出，通过互联网来进行预订生活产品的必要性。

旅游业作为经济社会的重要组成部分,在一个国家和地区一直占有十分重要的地位和作用。旅游业不仅在城市发展和经济社会发展中起到了积极的推动作用，而且也加快了酒店的发展和竞争。越来越多的酒店出现在大众的视野，然而对用户而言，根据自身的需求去选择酒店也成为了现代的潮流所向。

本文将详细地描述从数据的爬取、清洗到推荐系统的应用，让用户在最短的时间找到其感兴趣的酒店。

# 国内外研究状况

旅游电子商务于上世纪90年代初开始出现在欧美等西方发达国家。随着互联网技术的不断发展，旅游电子商务的应用模式和理论体系不断完善，给社会、企业和个人带来重大的影响。近几年，网上旅游份额增长迅猛。艾瑞咨询研究发现，在2009年，进行过在线预定的用户达2750万，同比增长95.0%，在线旅游预定市场规模达38.9亿元，同比增长32.3%。按照这种增长速度，到2013年，进行在线预定的用户人数将达到16800万。从营收结构上看，网上旅游预定市场主要以网上酒店预定业务为主，2009年网上酒店预定市场规模占整个在线预定市场规模的45.4%[ IResearch. 2009-2010年中国在线旅行预订行业发展告[R/OL]. http://www.iresearch.cn.]。

信息技术和电子商务对于各行各业的重要性已经不言而喻了，酒店行业也不例外。传统的企业在拥有自己的网站后，将更多的精力逐渐投入到网站功能的优化上，提高用户网络预定体验。一些在线代理商，如携程、艺龙等，则提供用户对酒店的评论和打分，借助 SNS的影响，扩大网站和酒店业务的影响。酒店提供的是一种体验式的服务。因此，消费者喜欢将自己的消费体验写在论坛、旅游博客中，而这些信息就是旅客的真实体验，与体验相关的这些评论的可信度较高，基本能够反映商家的服务质量，对那些潜在的预订者来说，这些评论具有很好的参考价值。相关研究[ Fact sheet[OL].2006. http://www.tripadvisor.com/pages/factsheet.html]表明，大部分在线酒店的访问者在预定酒店之前都会参考网站上的评论信息。在这些访问者中，84%的人在选择酒店时都会受到他们所看到的评论信息的影响。由此可见，酒店电子商务的推荐服务应该更加重视用户的评论和体验。然而遗憾的是，国内的酒店电子商务网站主要是通过酒店基本信息(包括星级、房型、价格等) 的介绍、酒店图片和酒店的综合排名等信息将酒店推向网民和客户，无法根据用户的个人体验和喜好向用户推荐满足其兴趣的酒店，也谈不上是个性化的服务。因此，研究用户的个性化体验打分，结合酒店的特征，向用户提供个性化的推荐，从而提高浏览者的对推荐信息的满意度。

# 论文结构

本文的安排组织如下：

本章主要介绍了推荐系统和国内以及国外的研究状况，说明我们这个推荐系统应用到酒店电子商务领域的重要意义，同时表明这个推荐系统的实现的价值和意义有了明确的研究目标。

第2章相关技术阐述，主要介绍的相关技术，如推荐系统、Hadoop平台、python爬虫等。首先介绍爬虫对酒店数据的爬取，其次到Hadoop平台去处理数据的具体操作，最后介绍了推荐系统的建立。

第3章介绍Hadoop集群中的设计方案，以及对数据的处理流程。

第4章介绍推荐平台的具体设计和实现方法。

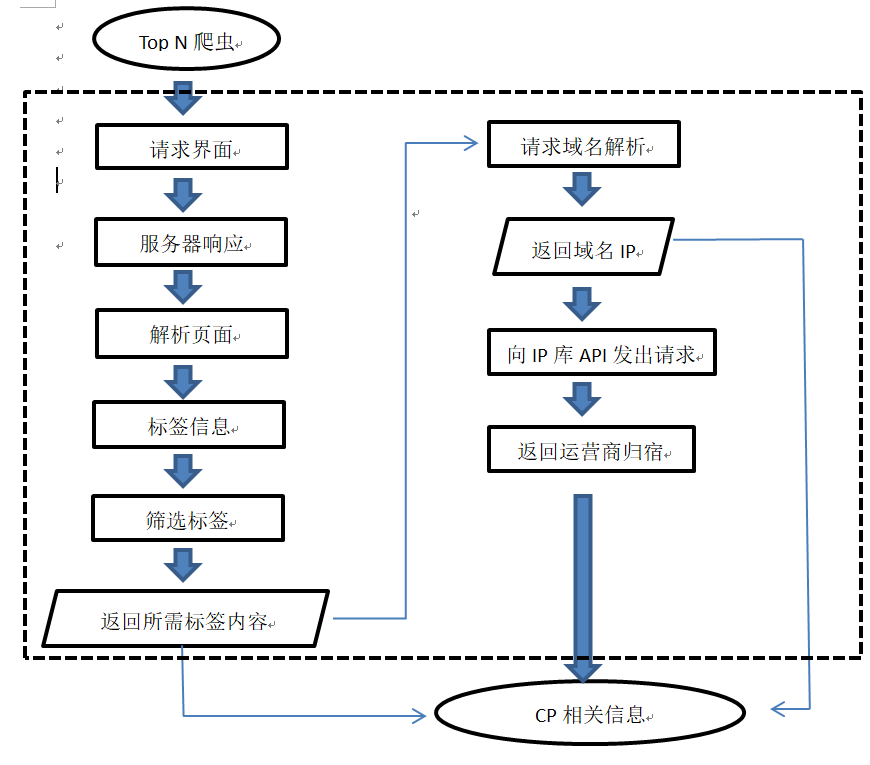
最后第5章总结全文，总结论文的主要工作，分析系统主要存在的问题，探索未来的研究方向。

**第二章 相关技术介绍**

# 2.1 Python爬虫技术介绍

随着我国互联网科技、大数据、云计算的迅猛发展的形势下，众多的行业产生的数据量随之变得越来越大，数据的海量化、开放化与事实化已成为现在数据挖掘和处理的主要特征。Python作为一门面向对象的计算机编程语言，其中存在大量的通用库和开放API，其可以满足更多网络中对数据的挖掘、处理分析需求，能够实现海量数据的查询、储存与展示。

Python语言是一门解释性编程语言，要想利用Python语言来开发网络爬虫就需要发挥出python的语言优势。Python爬虫工具的分为3个模块，分别为网站爬取、网站分析、数据储存，如图1所示。



**图1 网站爬虫程序流程**

python爬虫对网站中的数据进行抓取实现，其主要包括数据检索、数据抓取、数据分析等步骤，python为爬虫提供了大量丰富的库，其中python爬虫最为常用的就是Requests库、Beautiful Soup库以及lxml库这三大库。爬虫需要网络连接这个过程，也就是模拟计算机向服务器发送request请求和接收服务器返回的response内容进行识别，从而提取其中所需要的信息。

# 2.2 Hadoop平台介绍

Hadoop是由Apache所开发的开源分布式系统基础架构。可使用大量普通计算机或者服务器组成Hadoop集群。将计算机的储存能力和运算能力整合到一起，能够计算和储存更大规模的数据。Hadoop的MapReduce的分布式计算可以更快的实现数据的大批量处理和计算。Hadoop是由HDFS（Hadoop Distributed File System）和分布式计算框架MapReduce这两个核心组成。随着时间的推移，现在的Hadoop早已发展成为一个生态系统，如图2所示。

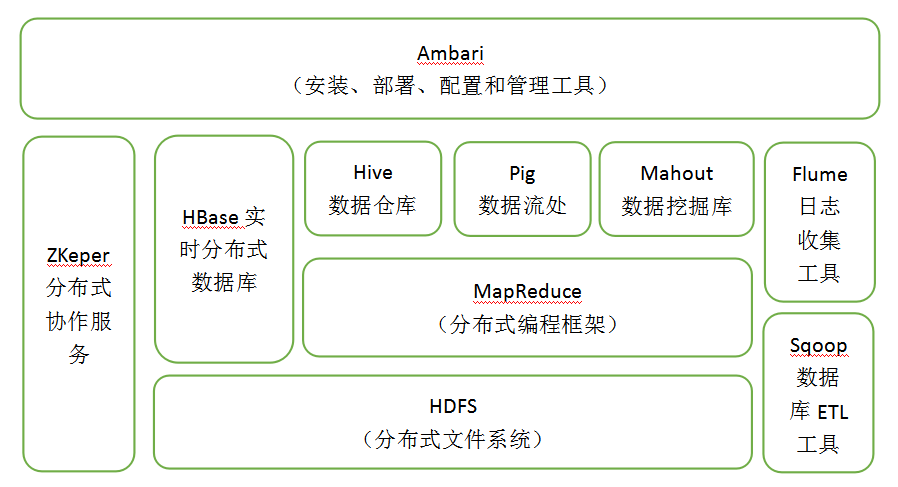


图2 Hadoop生态系统

HDFS采用主从结构模型，由NameNode和众多DataNode组成。NameNode为主节点，负责HDFS的命名空间管理和与客户端的交互；DataNode为从节点，负责分配数据的存储。HDFS为分块数据存储，也就是先切块，后存储。MapReduce是一个基于集群的高性能的分布式计算框架，它可以将一个很庞大的任务分发给手下的很多计算机或者服务器去进行高性能的计算。

HDFS和MapReduce是Hadoop平台的重要核心。HDFS实现了分布式集群上的文件系统管理，MapReduce实现了高性能分布式计算。HDFS为MapReduce提供相应的文件和储存支持，而MapReduce可以进行高效的任务分配、处理和监控等。它们之间的互相配合，共同实现Hadoop平台的分布式存储和计算的主要任务。

# 2.3 Spark的数据处理

Spark称之为第三代核心大数据处理框架，其集成第一代大数据框架Hadoop和第二代大数据框架Storm于一身，实现了基于内存的迭代式计算，在计算速度和吞吐量大小上远远比前两代框架厉害很多。Spark是用Scala语言编程开发的