信息科学技术专著丛书

知识驱动的Web查询处理技术

王芳 著

KNOWLEDGE-DRIVEN WEB QUERY PROCESSING TECHNOLOGY

非外借

北京邮电大学出版社
www.buptpress.com

信息科学技术专著丛书

知识驱动的 Web 查询处理技术

王 芳 著

北京邮电大学出版社
www.buptpress.com

内 容 简 介

Web查询处理旨在辅助信息检索系统更好地理解用户的信息需求与查询意图，该研究对于提高检索性能和改善用户查询体验具有重要作用，目前已成为信息检索领域最关注的问题之一。

本书关注于如何利用知识辅助查询处理：第1章简要概述Web查询处理的功能、主要研究内容及研究现状；第2章介绍了基于概念知识的Web查询分类，提出基于概念的短文本表示和相似度计算方法；第3章介绍了基于概念知识的Web查询理解，重在利用分类学知识从细粒度上识别用户查询意图；第4章介绍了CQA问题查询的命名实体消歧，重在挖掘先验知识辅助实体消歧；第5章介绍了基于大规模实体网络的相关查询推荐，利用实体相关性辅助单实体查询推荐。

本书适用于从事互联网搜索、文本理解、自然语言处理等研究方向的科研和技术开发人员，以及对大数据、人工智能、搜索引擎等技术感兴趣的读者。

图书在版编目(CIP)数据

知识驱动的Web查询处理技术 / 王芳著. -- 北京 : 北京邮电大学出版社，2022.7
ISBN 978-7-5635-6666-2

Ⅰ.①知… Ⅱ.①王… Ⅲ.①数据检索－研究 Ⅳ.①G254.926

中国版本图书馆CIP数据核字(2022)第103821号

策划编辑：马晓仟　　**责任编辑**：满志文　　**责任校对**：张会良　　**封面设计**：七星博纳

出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路10号
邮政编码：100876
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：唐山玺诚印务有限公司
开　　本：720 mm×1 000 mm　1/16
印　　张：8.5
字　　数：162千字
版　　次：2022年7月第1版
印　　次：2022年7月第1次印刷

ISBN 978-7-5635-6666-2　　定　价：38.00元

前　言

Web 查询是指由用户提交给信息检索系统用以表达其检索需求的文本。常见的 Web 查询有搜索引擎中的关键字查询、社区问答系统(CQA)中的问题查询等。当前的信息检索系统主要采用基于关键字匹配的检索模式,导致 Web 查询通常较短,只由较少的词或短语(又称查询词项)组成,而且往往具有模糊性和歧义性。信息检索系统难以理解用户真正的查询意图,仅依赖于基于查询词项的关键词匹配技术难以达到理想的检索效果。Web 查询处理旨在辅助信息检索系统更好地理解用户的信息需求与查询意图,该研究对于提高检索性能和改善用户查询体验具有重要作用,目前已成为信息检索领域最关注的问题之一。

近年来,随着大规模知识库的出现,如 Wikipedia、Freebase、Probase 等,越来越多的研究关注于如何利用知识辅助计算机理解用户查询需求。本书在已有工作的基础上,重点开展了基于知识库的查询分类、查询语义理解、查询实体消歧和查询推荐等方面的研究工作,具体包括以下四个方面。

(1) 基于概念知识的查询分类技术

现有的查询分类方法多利用词袋模型表示查询文本,常常受到字面不匹配的困扰。针对词袋模型在查询短文本表示中存在的诸多不足之处,本书提出了一种基于概念的查询表示方式,并在此基础上提出了一种新的查询分类框架。首先,利用分类学知识库,为每一个预定义类别学习一个概念模型,用以表示每个类别典型的概念信息;其次,提出了一种改进的查询短文本概念化(Conceptualization)方法,将给定查询短文本映射到一组相关概念中;最后,基于相同的概念空间,提出了一种概念相似度计算方法,依此进行短文本分类。实验表明,该分类框架在查询分类任务中四个类别的平均准确率高达 90.3%。同时利用概念信息,该框架在多样化排序中也取得了不错的效果。

(2) 基于概念知识的查询语义理解技术

理解用户 Web 查询背后真正的搜索意图是信息检索领域研究极具挑战性的

热点问题之一。其中，查询意图分类已经被广泛研究。本书进一步从查询修饰词-中心词的角度来理解查询。例如，给定查询“popular iphone 5 smart cover”，不同于使用粗粒度的意图类别（例如，电子产品），本书从细粒度查询意图识别的角度出发，旨在识别出“smart cover”是查询中心词，“iphone 5”是描述“smart cover”的修饰词，从而帮助搜索引擎更精准地查找意图相关的产品信息，即手机壳。该功能可以帮助搜索引擎精准获取相关搜索结果，同时对于广告匹配、查询推荐等应用也具有重要意义。本书提出了一种无监督修饰词-中心词检测方法。首先，从搜索日志中挖掘大量的实体层级的修饰词-中心词对。其次，设计了一个实体概念化方法，将实体层级的修饰词-中心词对泛化到概念级。再次，通过出现频次加权得到简洁、准确的修饰词-中心词概念模式，该过程旨在提升修饰词-中心词识别的泛化能力。最后，本书通过实验验证了所提方法的有效性。

（3）基于百科知识的查询命名实体消歧技术

命名实体消歧（Named-Entity-Disambiguation，NED）旨在将文本中具有多个含义的实体指称链接到知识库中具有明确含义的唯一实体。该技术在信息抽取、信息检索、机器翻译等领域有着重要的应用价值。当前研究工作较多地关注于长文本实体消歧，有关社区问答中问题查询的命名实体消歧研究较少。对问题查询进行命名实体消歧不仅有利于社区问答中的知识挖掘，还对问题检索、问题推荐以及问题路由等社区问答系统中的典型应用有帮助。其挑战性主要在于问题中的上下文信息较少、缺乏标注数据以及社区问答用户用语和知识库中的实体描述用语之间存在较大差异。为解决上述问题，本书提出了一种基于话题模型的问题命名实体消歧方法。具体而言，从知识库和社区问答元数据中挖掘问题与类别之间、实体与类别之间以及词项与实体之间存在的三方面先验知识，并将这些先验知识作为 Dirichlet 先验的超参数融入模型训练中，对问题生成过程进行弱监督。通过这种方式，所提方法无须人工标注就能够充分利用知识库和社区问答系统中的信息来丰富问题短文本，并在社区问答系统和知识库两种用语之间建立联系。

（4）基于大规模实体网络的查询推荐技术

Web 查询中经常会出现命名实体。本书将仅由一个实体构成的查询称为实体查询。实体查询作为 Web 查询的一个重要组成部分，在产品搜索、图片搜索等垂直搜索引擎中很常见。由于实体查询的长度很短，准确捕捉用户的查询意图非常困难。面向实体查询的相关实体推荐，旨在向用户推荐与原查询实体在不同方面或话题下相关的用户感兴趣的不同实体，对于引导用户查询意图、启发用户点击兴趣具有重要作用。已有的查询推荐技术研究大多以查询日志和查询相似文档为数据源，采用基于查询词项和点击的方法获取相关查询，较多地关注推荐内容的相

关性，而对于引导和启发用户的点击兴趣方面研究较少。此外，对于实体查询而言，从这些数据源中获取与原实体查询相关的实体非常困难。本书聚焦于相关实体推荐，利用网络中海量的实体描述页面提取大量相关实体，以此作为新的推荐数据源进行相关实体推荐。在相关实体提取过程中，本书利用实体之间的描述与被描述关系构建了一个大规模的相关实体网络。基于此相关实体网络，本书采用多种相关度计算方法对相关实体排序，比如基于共近邻实体和共近邻概念的语义相似度以及基于图结构的链接分析技术，以此用作相关实体推荐。实验表明，本书所构建的相关实体网络能够提供高质量的相关实体，基于大规模相关实体网络的相关实体推荐方法在准确率和新颖性方面均取得了不错的推荐效果。

以上四个方面的研究工作由知识驱动，分别针对普通的搜索查询(1)和(2)、长度略长的问题查询(3)和长度很短的实体查询(4)开展研究，重点阐述了如何利用知识辅助查询处理。自 2011 年起，作者在北京航空航天大学攻读博士学位，研究方向包括数据挖掘和自然语言处理；2012 年至 2015 年先后在微软亚洲研究院数据库组、机器学习组和自然语言处理组全职实习，从事该领域的算法研究和技术开发；自 2017 年入职北京石油化工学院以来继续该方向的研究工作。在此衷心感谢一路走来的所有同窗、老师、朋友和同事！本书也参阅了大量的国内外资料，未能一一列出，借此向这些著作和文献资料的作者表示衷心的感谢！

本书还得到了 2019 年北京市委组织部青年骨干个人项目(2018000020124G089)、2020 年北京市教委科技计划一般项目(KM202010017011)、2021 年北京市石油化工学院校级教育教学改革与研究重点项目(ZDFSGG202103001、ZDKCSZ202103002、ZD202103001)、北京市科学技术协会 2021—2023 年度青年人才托举工程项目以及北京石油化工学院交叉科研探索项目(BIPTCSF-006)的资助，在此一并感谢。

最后，感谢北京邮电大学出版社给予的大力支持。

尽管作者在本书撰写过程中耗费了很多精力，但由于水平有限，不足之处在所难免，恳请广大读者批评指正。

王　芳

2022 年 1 月

目　录

第 1 章

Web 查询处理概述

近年来，随着计算机技术的发展和互联网的普及，Web 上的资源以指数级迅速增长。根据中国互联网中心第 48 次《中国互联网络发展状况统计报告》，截至 2021 年 6 月，我国网民规模达 10.11 亿，较 2020 年 12 月增长 2 175 万，互联网普及率达 71.6%。十亿用户接入互联网，形成了全球最为庞大、生机勃勃的数字社会。超过 10 类互联网应用的用户规模在 5 亿以上，人均每周上网时长达到 26.9 个小时，互联网应用塑造了全新的生活方式和社会形态。互联网中蕴含着海量的数据资源，除海量网页以外，还有许多其他种类丰富的资源，包括图片、视频、用户自生内容(博客、微博、评论等)、开放链接数据和大规模知识库等。互联网日益成为人们学习、工作、生活的新空间，日益成为人们获取信息与公共服务的新平台。从互联网海量信息中准确快速地获取需要的信息成为是人们的客观需求。在此背景下，基于信息检索技术的搜索引擎应运而生，旨在帮助用户快速查找所需要的信息。

自 20 世纪 90 年代互联网搜索引擎出现以来，互联网搜索已经成为人们日常生活必不可少的一部分。特别是进入移动互联网时代。越来越多的人通过手机、平板电脑等移动设备接入互联网。搜索引擎正向精准化、智能化、个性化的方向发展。不断变化升级的计算机端搜索和移动搜索的背后，实际上是用户需求和市场格局的变迁。例如，随着人们更多地使用手机等移动设备上网，使用移动搜索引擎占用的也大都是零碎的时间，比如地铁上、客厅里、工作间隙等。在碎片化的时间里，用户需求指向更明确，他们没有耐心去一页页翻找信息，也很难忍受长篇大论，这就要求搜索引擎要更快更精准地给出用户所需要的信息。

1.1 搜索引擎工作原理

经典的商业网络搜索引擎，比如由百度、谷歌、Bing(必应)等，为提供全网搜索服务，需要处理数十亿甚至数万亿的网络文档，并且文档数量一直在持续更新，需要 PB 级的数据存储空间，并确保通过数十亿用户查询来满足搜索引擎用户的需求。一般来说，搜索引擎需要准确地了解搜索查询，然后根据用户输入的查询有效从海量文档中找到相关结果并对结果进行排序，最终将排序后的结果呈现给用户。如图 1-1 所示[①]，上述检索过程包含线上和线下两部分。

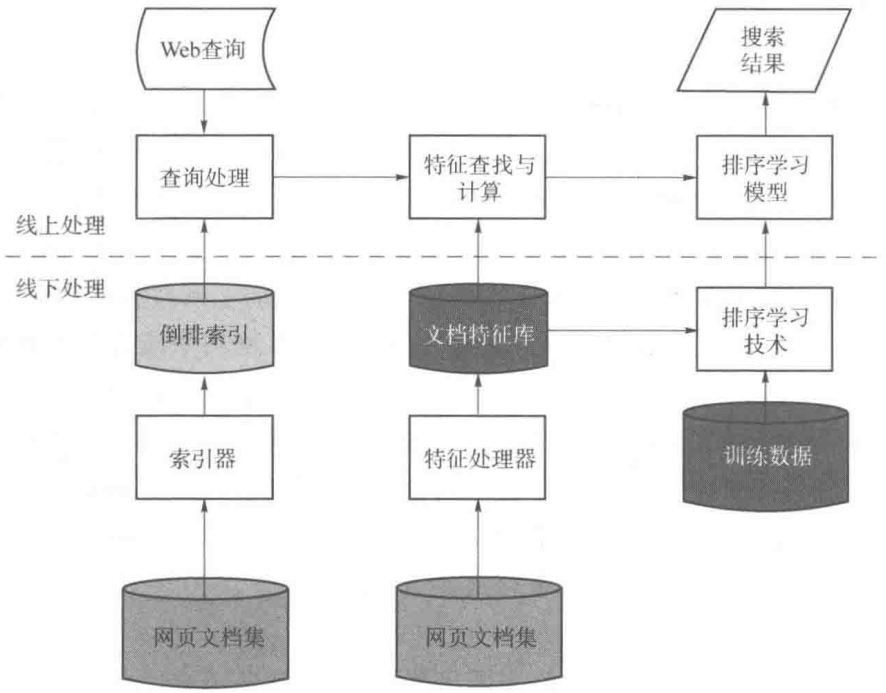

图 1-1 搜索引擎工作基本原理示意图

(1) 线下部分：主要完成网页文档信息采集、信息索引和为实时信息检索做准备。

1) 网页数据采集

互联网上的信息存储在无数个服务器上，任何搜索引擎想要满足用户的搜索需求，首先需要把网页存在自己本地的服务器上，这靠的就是网络爬虫。网络爬虫

① 本章此处仅对搜索引擎工作原理进行基本介绍，不展开技术细节说明。

不停地向各种网站发送网页请求，将所得到的网页存储到搜索引擎服务器，构成网页文档集。

2）建立倒排索引

面对海量文档集，如何根据用户查询找到相关的网页文档，是搜索引擎要解决的第一个问题，这靠的就是倒排索引。倒排索引类似字典，是一张单词索引表，它记录了单词在多少文档中出现、分别是哪些文档、每个文档出现了多少次、分别出现在什么位置等信息。这样在检索文档中，只需将用户查询分词，根据查询词检索对应出现的文档，就可以很快捷地找到与搜索查询相关的网页文档。

3）文档特征处理

文档处理在建立索引和结果排序模型训练中具有重要意义。由于网络上的文档类型多种多样，针对每一种格式的文档都要有一个对应的解析器，用于提取有用内容，解析出干净的文档文本后，还需要用到自然语言处理技术对文档进行处理，例如英文文档的分词、词干提取、词性识别、创建 n-gram 模型等。经过处理后的文档，需要进一步进行特征提取，用于训练搜索结果排序模型。常见的文档特征包括表征网页链接关系的 PageRank 值、表征词在文档中的重要性指标 TFIDF、编辑距离等。

4）排序学习

对搜索结果进行排序，其目标是设计并应用一种方法来自动地从训练数据学习训练出一个函数，这个函数的功能是可以依据在某个特定的应用中定义的相关性、用户偏爱或在特定应用领域中的重要性来将对象（如文件、网页等）进行排序。这一训练学习过程称为排序学习，是信息检索和机器学习领域研究的热点问题。给定训练数据，设计学习模型，从而为线上实时对搜索结果进行排序做准备。

（2）线上部分：主要负责实时对用户查询进行处理、特征查找与计算以及结果排序。

1）查询处理

当收到用户的搜索查询时，搜索引擎需要根据查询中的词项，选择要返回给用户的文档。在严格的解释中，查询将精确定义检索文档必须/不必须包含的词。但实际上，用户的搜索查询往往存在语法和拼写等错误，甚至用户自己难以用准确的查询词表达其查询意图。为此需要先对查询进行处理，从而更好地进行信息检索和结果排序。

2）特征查找与计算

此步骤类似上述离线部分的第 3）步。将用户搜索查询看成一个短文本，为进行结果排序，需要对其进行预处理和特征提取。

3) 结果排序

此步骤在 Web 查询和相关文档特征提取的基础上，利用离线阶段训练好的排序学习模型，对检索到的相关文档进行排序，排序后按顺序将搜索结果返回给搜索用户。

1.2 Web 查询处理简介

由上述搜索引擎基本原理可知，当前的信息检索系统主要采用关键字匹配的检索模式。用户通过查询接口向信息检索系统提交一段文本作为查询，信息检索系统则根据用户查询给出检索结果。Web 查询通常是一些短文本片段，这些查询具有自然语言的特点，但相比自然语言，其表达方式更为随意，主要特点表现在以下三个方面。

(1) 查询的长度通常较短，只由较少的词或短语(通常称为查询词项)组成，根据哈尔滨工业大学自然语言处理实验室基于搜狗查询日志的分析，中文查询的平均长度为 3.4 个词，根据 Silverstein 等人的调查，英文查询的平均长度为 2.3 个词。

(2) 查询词项多为表示具体意义的实词。根据文献[2]中对英文查询日志的调查，70%的查询中包含命名实体，表明了用户经常采用命名实体作为检索的查询词项。查询实体之间常常缺少衔接的虚词和一些表示语法关系的上下文词汇，例如用户想要查询“iPhone 6”的保护套，通常会搜索“cover iphone 6”。

(3) 查询往往具有模糊性和歧义性。例如查询“apple ipad”中的“apple”是一个具有歧义的词项，可以指代电子产品品牌也可以指代一种水果。如果搜索引擎不能正确识别该查询的语义，很可能错误地返回有关水果的检索结果。

Web 查询处理是搜索引擎进行信息检索的第一步，也是用户和搜索引擎进行交互的关键步骤。有效处理查询，对于提高搜索引擎信息检索准确性、提升用户搜索体验具有重要意义。同时也与当前搜索引擎的发展紧密相关，主要体现在以下两个方面。

(1) 向用户提供准确的查询信息一直是搜索引擎的研究热点。由于查询的以上特点，当前基于关键字匹配模式的搜索引擎无法达到满意的查询效果。例如，当用户输入查询“angry bird iphone”时，错误地返回大量有关“iphone”甚至是“bird”的结果将大大降低用户查询体验。理解查询短文本背后用户真正的信息需求，有助于搜索引擎为用户提供准确的检索结果。此外，搜索引擎中存储着越来越多的结构化和半结构化的数据，在这些结构化数据资源上进行的检索能够得到更直接和准确的结果。例如，如果搜索引擎的后台数据库中存储了电影放映信息数据，当

用户查询“功夫熊猫上映时间”时，搜索引擎能够直接将上映时间信息准确地提供给用户。这种基于结构化数据的信息检索在很大程度上依赖于搜索引擎对用户查询的正确理解与分析。例如，需要对查询意图类别加以识别，进而判断查询是否可以通过结构化数据资源回答。此外，还需要分析查询的检索模式，以便和后台的结构化数据相匹配。

（2）从提供信息到提供服务的转变是当前搜索引擎的一个发展趋势。在基于服务平台的搜索引擎中，用户希望通过搜索引擎不仅能够从互联网上获得网页信息，还能够直接得到所需的服务。从搜索引擎的角度上，也希望可以提供给用户更广泛的互联网入口。当前的搜索引擎致力于成为提供互联网上信息、资源（如音频、视频、图像等）以及交互应用（如地图、购物、本地生活服务、新闻、社交等）的服务平台。这种基于服务平台的搜索引擎，迫切需要准确理解用户的查询需求，向用户提供更加个性化、场景化的精准信息搜索服务。例如，需要对用户查询的检索范围进行限定，明确用户所需的服务类型。举例来说，对于查询“长津湖下载”和“长津湖放上映时间”，需要识别出用户查询的不同需求，如电影下载和上映时间查询，使得搜索引擎能够准确地将该查询分配给相应的应用或内容资源提供商进行处理，从而返回满足用户需求的信息服务。

总而言之，Web 查询处理技术研究对于提高信息检索性能和改善用户检索体验具有重要作用。图 1-2 所示为 Web 查询处理示意图。常见的查询处理包括查询自动补全、查询分类、查询语义理解、查询修正与查询重写、查询推荐等。

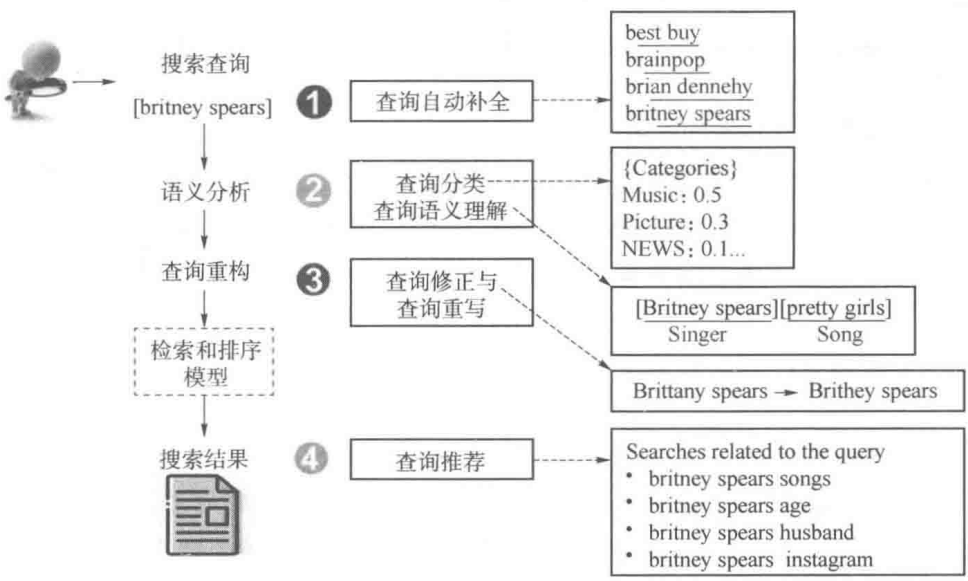

图 1-2　Web 查询处理示意图

(1) 查询自动补全

查询自动补全是指当用户输入查询词时,根据用户当前输入的查询给出完整的查询推荐,并显示在搜索框适当位置作为搜索提示。如图 1-2 所示,用户想要查询"britney spears",从用户输入"b"到"brit"时,搜索引擎根据用户输入的字母,给出最相关的完成查询提示。一方面查询补全可以帮助用户减少键盘输入,增加用户体验;另一方面当用户也不明确自己想要查询什么的时候,通过查询补全可以给出相关热门搜索推荐,帮助用户明确查询意图。

(2) 查询分类

当用户输入 Web 查询点击搜索按钮后,查询处理的主要任务就是语义分析,旨在准确理解用户查询背后的搜索意图。查询分类是语义分析的经典任务之一,目标是将用户提交的搜索查询自动分类到预先设定的类别。如图 1-2 所示,给定查询"britney spears pretty girls",可以将其分类到音乐、图片和新闻等类别,属于各类别的概率分别是 50%、30%和 10%等。查询分类有助于在搜索引擎根据类别对搜索结果排序,从而给出更好的排序结果,提升用户体验和搜索点击率。

(3) 查询语义理解

查询语义理解旨在理解查询中蕴含的语义信息,相对于查询分类任务所识别的查询类别信息,语义理解所得到的语义信息更加细化。如图 1-2 所示,给定查询"britney spears pretty girls",语义理解可以识别出"britney spears"是一名歌手,"pretty girls"是一首歌曲。这对于搜索结果排序具有更精确的指导意义。甚至可以进一步推断用户想要查询的是"britney spears"演唱的"pretty girls"这首歌,查询中心词是"pretty girls","britney spears"是修饰词。基于此,搜索引擎可以给出精确的查询结果。

(4) 查询修正与查询重写

在正式开始信息检索之前,查询处理的另一项重要任务是将原有查询进行改造重写,通过查询扩展、拼写修正、查询重写等方式进一步完善查询,其目的是最大限度地增加相关文档的检索覆盖率,避免由于单词本身和文档正文单词之间的字面差别导致的文档漏检。如图 1-2 所示,假设用户输入的查询是"brittany spears",在这个处理步骤中可以将其重写为"britney spears",用以查找到更多相关的网页文档。

(5) 查询推荐

上述查询处理步骤都是在搜索引擎进行信息检索之前完成的。当搜索引擎给出排序结果后,同时给出相关的查询推荐。这一步骤同样很关键,尤其当用户对本次搜索结果不甚满意的时候,相关查询推荐,对于进一步引导用于明确查询需求至关重要。此外,查询推荐可以激发用户的查询需求,特别是个性化查询推荐,可以

在建模用户兴趣的基础上，推荐用户感兴趣的相关查询，增加用户体验，让用户“搜”罢不能。如图 1-2 所示，用户当前查询为“briteny spears”，查询推荐包括“briteny spears”的歌曲、“briteny spears”的年龄、“briteny spears”的丈夫及其 instagram 社交账号等。

本书所述的 Web 查询处理关键技术，聚焦于对查询的语义分析，主要包括查询分类、查询意图理解、查询实体消歧、查询推荐和关键词提取等内容，不涉及查询处理中的基础技术问题，例如分词、词干提取、词形还原以及拼写修正和查询补全等。

1.3　相关研究现状

查询处理相关技术已经成为信息检索领域最关注的问题之一，已有大量相关研究工作。本节针对所研究的主要内容，简要介绍查询分类、查询意图理解、查询消歧和查询推荐四方面的相关研究工作，与本书内容相关的更具体的研究工作将分别在行文中给出介绍。

1.3.1　查询分类

Web 查询分类不仅可用于限定用户查询的检索范围，明确用户的搜索意图，对于提高搜索引擎的搜索质量也具有重要的意义。近年来，查询分类已经成为信息检索领域的一个研究热点。本节首先介绍查询常用的类别设置，随后详细介绍已有的分类方法。

(1) 类别设置

目前，查询分类并没有统一的分类体系。Broder 等人[1]通过分析查询日志并对用户进行调查，将查询意图分为导航类、信息类和事务类。

① 导航类：该类查询的意图是想找到某一特定的网站，例如“新浪博客首页”“百度”“网易 NBA”等。

② 信息类：该类查询的意图是想找到一些存在互联网上的静态的信息，例如“汽车”“人民的名义”等。

③ 事务类：该类查询的意图是寻找一些能够进一步交互的网站，比如说“iPhone 网购”“QQ 去广告版”“人民的名义在线观看”等。

Daniel 等人[2]在 Broder 的研究基础上，进一步细化了查询意图类别。他们的分类体系中包含导航类、信息类和资源类，其中导航类包含 5 个子类、资源类包含 4 个子类。