信息科学技术学术著作丛书

查询推荐理论与方法

蔡　飞　陈洪辉　蒋丹阳　陈皖玉　著

科学出版社

信息科学技术学术著作丛书

查询推荐理论与方法

蔡　飞　陈洪辉　蒋丹阳　陈皖玉　著

科学出版社

北　京

内 容 简 介

本书较全面地介绍信息检索中查询推荐理论与方法，描述查询推荐的研究背景、模型概述、实验框架和实现方法。具体阐述前缀自适应和时间敏感的个性化查询推荐、基于同源查询词和语义相关性的查询推荐、多样化查询推荐和选择性个性化查询推荐等理论方法。

本书许多内容是作者近年来在信息检索领域的最新研究成果，具有较强的学术性和原创性。本书内容丰富、概念准确、叙述严谨、图文并茂，理论与实验相结合，可作为高等院校和科研院所计算机科学与技术、软件工程、计算机应用技术、信息系统工程等相关专业的高年级本科生或者研究生的参考书，也可供软件开发相关领域的研究人员借鉴和参考。

图书在版编目(CIP)数据

查询推荐理论与方法/蔡飞等著. —北京：科学出版社，2017

(信息科学技术学术著作丛书)

ISBN 978-7-03-051200-0

Ⅰ. 查… Ⅱ. 蔡… Ⅲ. 信息检索 Ⅳ. G254.9

中国版本图书馆 CIP 数据核字(2016)第 303224 号

责任编辑：魏英杰 / 责任校对：桂伟利

责任印制：张 倩 / 封面设计：陈 敬

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 1 月第 一 版 开本：720×1000 1/16

2017 年 1 月第一次印刷 印张：12 3/4

字数：260 000

定价：80.00 元

(如有印装质量问题，我社负责调换)

《信息科学技术学术著作丛书》序

21 世纪是信息科学技术发生深刻变革的时代，一场以网络科学、高性能计算和仿真、智能科学、计算思维为特征的信息科学革命正在兴起。信息科学技术正在逐步融入各个应用领域并与生物、纳米、认知等交织在一起，悄然改变着我们的生活方式。信息科学技术已经成为人类社会进步过程中发展最快、交叉渗透性最强、应用面最广的关键技术。

如何进一步推动我国信息科学技术的研究与发展；如何将信息技术发展的新理论、新方法与研究成果转化为社会发展的推动力；如何抓住信息技术深刻发展变革的机遇，提升我国自主创新和可持续发展的能力？这些问题的解答都离不开我国科技工作者和工程技术人员的求索和艰辛付出。为这些科技工作者和工程技术人员提供一个良好的出版环境和平台，将这些科技成就迅速转化为智力成果，将对我国信息科学技术的发展起到重要的推动作用。

《信息科学技术学术著作丛书》是科学出版社在广泛征求专家意见的基础上，经过长期考察、反复论证之后组织出版的。这套丛书旨在传播网络科学和未来网络技术，微电子、光电子和量子信息技术、超级计算机、软件和信息存储技术、数据知识化和基于知识处理的未来信息服务业、低成本信息化和用信息技术提升传统产业，智能与认知科学、生物信息学、社会信息学等前言交叉科学，信息科学基础理论，信息安全等几个未来信息科学技术重点发展领域的优秀科研成果。丛书力争起点高、内容新、导向性强，具有一定的原创性，体现出科学出版社“高层次、高质量、高水平”的特色和“严肃、严密、严格”的优良作风。

希望这套丛书的出版，能为我国信息科学技术的发展、创新和突破带来一些启迪和帮助。同时，欢迎广大读者提出好的建议，以促进和完善丛书的出版工作。

中国工程院院士
原中国科学院计算技术研究所所长

此为试读，需要完整PDF请访问：www.ertongbook.com

前　言

查询推荐是当今主流搜索引擎或者信息检索系统的一个主要功能，就是当用户使用搜索引擎或者信息检索系统时，在输入查询短语的过程中，检索系统根据用户输入的查询短语前缀，推荐给用户一组查询候选词供用户点击选择，从而帮助用户完成查询短语的构建。这些查询短语往往都是其他用户先前提交给检索系统的查询短语。目前，查询推荐方法主要依赖于检索系统搜集到的用户查询日志，来挖掘用户相关信息，预测用户查询意图，从而提供给用户一组其最有可能提交的查询短语列表。

在本书中，我们将介绍信息检索中查询推荐相关理论与实现方法，包括时效性敏感的个性化查询推荐方法、基于机器学习的查询推荐模型、多样化查询推荐方法和选择性的个性化查询推荐方法。通过大量的实验验证，本书提出的查询推荐方法合理有效，模型性能指标显著提高。我们相信，本书中的研究成果可以为搜索引擎的功能设计和检索系统的查询推荐优化提供帮助，辅助信息检索用户快速获取所需的信息，从而提高用户使用满意度。

本书共8章。第2章作为查询推荐方法的相关研究概述，可以使读者对当前研究现状有初步的了解。第3章介绍研究章节中实验部分的测试数据集，以及算法评估度量指标，同时简单介绍用作实验比较的若干查询推荐传统方法。第4章～第7章可单独阅读，因为这些章节的主要内容相对于其他章节是独立的，分别介绍不同的查询推荐算法模型。最后，阅读第1章和第8章可以得到本书各个研究内容的一个小结，而且为本章中前面提出的问题提供深刻见解。

本书内容是国防科学技术大学信息系统与管理学院信息系统工程重点实验室众多科研人员多年学习、研究沉淀的成果。第 1 章、第 2 章和第 8 章由蔡飞撰写，第 3 章由陈洪辉撰写，第 4 章和第 5 章由蒋丹阳撰写，第 6 章和第 7 章由陈皖玉撰写。陈洪辉负责全书的内容组织与统稿。罗雪山教授、刘俊先教授、罗爱民教授、舒振副研究员、陈涛讲师等对本书的撰写提供了指导意见，在此对他们的辛勤工作和热心帮助表示衷心的感谢。本书的出版得到了装备预先研究项目“XX 主动推荐技术”(315100103)、“XX 能力评估技术”(315010201)、“XX 结构框架及工具”(315010103)、“XX 设计与优化技术”(414050103)和装备预研基金项目“XX 概念体系和关键技术”(6141B08010101)等项目的资助，在此一并表示感谢。

查询推荐理论与方法是当前信息检索领域处于科学前沿的课题之一，相关的理论和技术还在发展之中，新的查询推荐思想、理论和方法技术还在不断完善和验证中。限于作者水平，书中不妥之处在所难免，恳请读者批评指正，共同推动查询推荐理论和方法研究的进步和发展。

作　者

2016 年 10 月于长沙

目　录

此为试读,需要完整PDF请访问: www.ertongbook.com

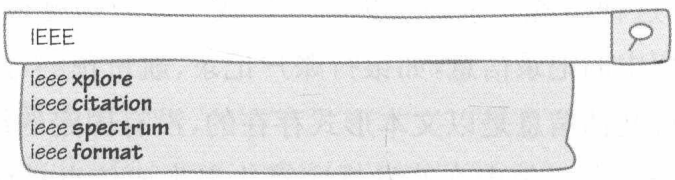
第1章 绪 论

信息检索(information retrieval,IR)是指信息按照一定的方式组织起来,并根据用户的需要找出能解决用户信息需求的过程和技术。可见信息检索的主要目的是解决用户的信息需求。一般信息检索活动主要包括用户向搜索系统提交查询词汇、检索系统计算文档和查询的相关度、检索系统返回与查询相关的文档。20世纪50年代以来,信息检索领域研究的主要检索对象是文本文档,如网页、邮件、学术论文、书籍和新闻报道等[1]。这些典型的文档由特定的格式来存储标题、作者、日期和摘要等,这种结构要素通常称为属性或字段。信息检索中的文档信息与数据库中的记录信息,如银行账户记录、航班预约等的重要区别在于绝大部分文档信息是以文本形式存在的,没有固定的结构属性[1]。信息检索的研究主要集中在开发排序算法来生成信息排序列表,回应用户的查询,从而满足他们的需求。

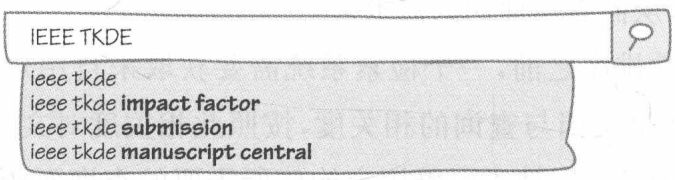
在对文件排序之前,一个检索系统需要获取来自用户的查询。然后,系统估计出文档与查询的相关度,按照此相关度对文件进行排序。查询构建是为了帮助搜索引擎用户构建所需要的查询而产生的。信息检索系统主要包括一个文本库,用于计算权重、扩展与查询构建活动有关的内容。查询构建的主要范围涉及查询推荐、查询重写、查询转换等,目的是为了更好地描述用户的潜在查询意图。因此,查询构建的最终目标是提高返回给用户的排序文件的整体相关度。作为查询构建工作的一项任务,查询推荐(query auto completion)需要在用户仅仅输入一些前缀,如几个字符时就可以帮助用户构想出整个查询问题[2-4]。查询推荐的主要目的是预测用户的查询意图,从而减少用户输入查询的

字符数，减少用户的查询构建时间。随着即时搜索，如 Google Instant 的出现，正确的查询推荐变得非常重要，这决定了用户获取正确信息的时间，因为用户可以及时看到检索系统返回的相关结果[3]。查询推荐已经成为当下主流搜索引擎，如 Bing、Google、Yahoo!、Baidu 等，以及一些在线应用的显著功能，如网购、电子邮件服务等。在查询推荐的预计算系统中，与查询前缀匹配的查询候选词汇列表是预先生成并存储在一个高效数据结构中的，这样做的目的是便于快速查找。

如图 1.1 所示，在信息检索系统中，输入字符时系统会首先返回一组匹配的查询推荐列表，当用户继续输入字符时，查询推荐列表也在不断地更新。检索系统查询推荐功能的应用极大地改善了用户检索体验，对检索结果的满意度也产生了深远的影响，因此查询推荐被搜索引擎用户广泛采纳[3,4]。



(a) 查询前缀“IEEE”的查询推荐实例



(b) 查询前缀“IEEE TKDE”的查询推荐实例

图 1.1 检索系统查询推荐实例

在查询推荐问题上，一个直接有效的方法是从一段时间内的查询记录中提取查询及其前缀，统计各个查询短语的出现次数，即频率，再根据查询短语的频率进行排序[3-5]。这种方法假定当前和未来查询的频率与过去已知的查询频率保持一致。虽然这种方法得出的查询推荐结果在满意度上较好，但远达不到最佳状态，因为这种方法未能将时

间、趋势和特定用户的偏好等重要因素考虑在内，而这些信息通常会影响查询推荐排序的结果。

早前的工作[3,4]表明，时间序列分析技术可以用来对带有季节性意图的查询进行分类，并预测其未来的查询次数，这意味着可以将这些策略嵌入到基于频率的查询推荐和查询分类方法中。本书将继续调查查询频率的特性变化，并开发一种时效性的查询推荐方法。这种方法不仅仅是结合用户个人搜索记录形成的偏好模型，而是将短期记录（当前会话）和长期记录（先前历史）都考虑在内。提出的方法能够提升对用户查询意图的感知，能将用户最终提交的查询返回在查询推荐列表的靠前位置。在此基础上，扩展初始的查询推荐模型，我们通过学习优化，使得初始模型中来自查询频率和用户特征的贡献达到最优化，因此可以更好地为一些不常见的查询前缀提供查询推荐列表。

基于频率的查询推荐方法[3-5]在统计查询次数时遵循严格的查询匹配原则，使用这种方法，我们认为相似查询对基于查询频率的查询推荐方法也至关重要。因此，在对最初查询备选词排序时，可以将同源相似查询的频率也考虑进来。同源相似查询主要包括与原始查询具有相同的查询字，但顺序不同的查询，或者对原始查询备选词汇进行适当扩充的查询。此外，我们将查询字之间的语义相似度特征融入查询推荐方法中，认为用户在构建查询时倾向于将语义上有联系的词汇组合起来。因此，基于一组最新构建的特征，我们提出一个基于排序学习的查询推荐方法，通过对特征的训练，直接建立查询推荐的排序模型。实验表明，语义关联性和同源相似查询的特征非常重要，而且确实能带来查询推荐性能的提升。

本书工作的第3个重点是多样化查询推荐，这在目前还未被很好地研究。先前的研究工作主要围绕在反馈给用户一组查询推荐列表，着眼于将用户最可能想输入的查询放在顶部，但是却忽略了备选查询列表中的冗余度。因此，输入前缀的语义相关的查询经常会被一起返

回给用户，有可能导致有价值的查询推荐无法出现在有限长度的查询推荐列表中，进而导致非最大化的用户满意度。与多样化网页搜索结果[6-10]目的不同，多样化查询推荐的目的是既要将用户可能的查询推荐返回在查询推荐列表的靠前位置，同时又要减少查询推荐列表的冗余度。我们提出一个查询推荐的贪婪算法，这种算法是以备选查询词汇当前的搜索频率和在同一个会话中先前的查询意图为依据，来预测查询推荐排序。我们采用国际公认的指标来衡量查询推荐的准确率和查询推荐的多样化。

为了满足用户特定的信息需求，可以将用户的搜索历史和兴趣考虑进来[11]，建立个性化的用户查询推荐模型。个性化查询推荐方法的研究已经开展了一段时间，取得了较好的实验效果[3,12]。这类个性化查询推荐方法以一个固定的方式来个性化查询推荐列表。然而，我们认为个性化在不同检索环境中对查询推荐列表的排序作用并不一致，因此我们的关注点在于将个性化有选择地融入到查询推荐模型中。基于个性化查询推荐策略（考虑查询频率和个性化搜索意图），提出选择性的个性化查询推荐（SP-QAC）模型，来权衡查询频率和个性化搜索意图的贡献。具体而言，我们以一个回归模型为基础，在每种查询推荐测试中估计预测权衡参数，主要依据来自用户已经输入的查询前缀、点击的文件和先前的查询等，这些信息都会用于挖掘个性化在查询推荐模型中的权重。实验结果表明，个性化可以有选择地嵌入到一个传统查询推荐模型中，而不是固定地应用到查询推荐的框架中，这样可以进一步提高查询推荐的性能[13]。

1.1 研究概述与研究问题

推动本书研究的最终问题是，我们怎样提高信息检索中查询推荐（查询推荐）的性能。本书第 2 章对解决这个问题提出一些方法策略，

但是诸如如何更好地将时间、用户情景或语义等信息融入查询推荐方法中还有待研究。本书将主要介绍信息检索领域查询推荐研究的相关理论和技术方法。

首先,我们把重点放在如何结合查询推荐中时效性特点和用户个性化特性上。在之前的工作中,时效性查询推荐模型和用户个性化查询推荐方法已经分别得到发展。不论是时效性,还是个性化,这两类查询推荐方法都取得了较好的研究进展。在本书中,我们首先提出一种时效性查询推荐模型[14,15],也就是 λ-TS-QAC 模型和 λ^*-TS-QAC 模型,分别引入一个固定的权衡因子 λ 和一个最优权衡因子 λ^*,来控制预测查询频率时近期趋势和周期性信息的贡献度。考虑到查询搜索频率的季节性变化[4]和近期趋势[11]可用于预测查询短语的未来频率,我们试图了解这两部分是如何整合的,来提高基于时效性信息预测查询频率的准确性。对比多种预测模型得出结果,同时来回答以下几个研究问题。

问题 1:作为一个合理的测试,多种不同预测模型得到的查询频率预测准确性怎么样?

问题 2:提出的时效性查询推荐模型(λ-TS-QAC 和 λ^*-TS-QAC)性能相比当前最先进的时效性查询推荐方法的准确率如何?

在回答这两个研究问题的过程中,我们发现基于周期性和近期查询流行趋势的预测方法可以得出查询频率的正确预测,而且就预测指标平均绝对误差(mean absolute error,MAE)和对称平均百分比误差(symmetric mean absolute percentage error,SMAPE)来说,比其他基于预测模型都要好。基于各预测模型的查询频率,我们提出的时效性查询推荐模型的查询推荐准确率指标排序倒数均值(mean reciprocal rank,MRR)有较大幅度的提高。

在此基础上,我们提出一种混合查询推荐模型 λ^*-H-QAC,将时效性和个性化因素同时考虑进来,并与基于 n-gram 的混合查询推荐模型 λ^*-H_G-QAC 进行了性能比较。此外,提出 λ^*-H-QAC 模型的扩充模

型 λ^*-H'-QAC，通过最优化查询频率预测贡献和用户个性化特征贡献的权重，来处理长尾前缀，即非常见前缀的查询推荐。为了验证提出模型的有效性，我们回答以下问题。

问题 3：λ^*-H-QAC 模型的查询推荐准确率相比传统的时效性查询推荐方法，如 λ^*-TS-QAC 性能有何变化？

问题 4：相比基于 n-gram 的混合个性化查询推荐模型，λ^*-H-QAC 模型性能如何？

问题 5：λ^*-H-QAC 模型和 λ^*-H_G-QAC 模型的实验结果有何差异？

问题 6：在长尾前缀的查询推荐问题上，λ^*-H'-QAC 模型和 λ^*-H-QAC 模型性能有何差异？

实验结果表明，在整合了以用户为中心的搜索环境和时效性查询推荐模型后，我们的方案，即混合查询推荐方法大大提高了查询推荐的排序效果。

通过分析先前发展较为成熟的查询推荐技术，可以发现当下大多数查询推荐模型是用频率（即查询次数）给备选查询排序的，但是这类方法在统计查询次数时遵循严格的查询匹配原则。也就是说，忽略来自相似查询（查询字相同但顺序不同的查询或对原始查询进行扩充的查询）的贡献。由于相似查询往往表达出极其相似的搜索意图，而且目前的查询推荐方法经常忽视查询字之间的语义相关性。然而，用户在构建查询时倾向于将语意相关的字词结合起来。为了处理这一缺陷，在基于用户行为分析的排序学习查询推荐模型（L2R-U 模型）[12] 的基础上，我们提出一组基于排序学习查询推荐模型，将来源于预测频率、相似查询和语义相关性等特征分别引入询推荐模型。

具体而言，我们提出如下模型。

① L2R-UP 模型，挖掘查询推荐短语的观察频率和预测频率。

② L2R-UPH 模型，挖掘查询推荐短语的同源相似查询的观察频

此为试读，需要完整PDF请访问：www.ertongbook.com