学术文献引文推荐研究进展*

陈海华 孟睿 陆伟
武汉大学信息管理学院 武汉 430072 武汉大学信息检索与知识挖掘研究所 武汉 430072

摘要: [目的/意义] 学术文献引文推荐是指对于给定的学术文献,自动化地为其推荐合适的引文和参考文献。借助于引文推荐,用户可以在一定程度上提高撰写学术文献的效率,降低对重要相关文献的漏引。[方法/过程] 分析国内外引文推荐研究的最新进展,阐述引文推荐问题的演化过程,从局部引文推荐和全局引文推荐等方面对引文推荐进行梳理,重点归纳文档相似性、主题模型、翻译模型、协同过滤和混合推荐等 5 种引文推荐常用方法,并总结引文推荐常用数据集和测评方法。[结果/结论] 已有引文推荐研究的主要问题在于未考虑用户偏好的动态变化性及研究领域的综合性,在用户研究和实际应用方面仍有所欠缺; 未来引文推荐的研究可运用语义化表达方法和自然语言生成技术,从基于上下文的引文推荐和跨语言引文推荐等方面进行展开。

关键词: 引文推荐 引文推荐分类 引文推荐方法 引文上下文
分类号: G353.4
DOI: 10.13266/j.issn.0252-3116.2015.15.018

1 引言

引用相关研究成果是研究者在撰写学术文献时的重要环节,作者引用其他文献的原因和动机大致有以下 3 种[1-2]: ①从文献中获取写作灵感,借鉴其研究思路、研究方法或学术观点; ②在论证自己观点的时候引用其作为论据,使自己的观点更加充实饱满; ③在相关工作中将其作为最新研究进展加以描述。研究者在撰写学术文献时往往需要引用大量的参考文献来支撑自己的观点,并且不同学科之间所需的引文数量差异巨大[3],特别是一些相对较成熟的学科(如生物学等)有时甚至需要穷尽所有相关的参考文献,这必然会耗费研究者大量的精力。在学术文献数量飞速增长的当代,每天都有数以万计的学术成果被发表,据 R. M. May[4] 1997 年的统计结果,公开出版物的年增长速率为 3.7%,特别是在一些比较热门的研究领域,现在这个数字更为惊人[5]。如何迅速地在质量参差不齐的学术资源中找到合适的相关文献是科研人员面临的一大挑战。目前研究者们通常会借助于一些学术文献管理工具来组织相关参考文献,当下最为流行的文献管理工具有: LaTeX[6]、NoteExpress[7]、EndNote[8] 等。以 LaTeX 为例,其工作流程[9] 大致为: 研究者根据写作需求在网上检索合适的资源,通过仔细阅读手动筛选出自己所需的参考文献,再将参考文献与特定的引文句一一对应,这是一个反复斟酌的过程。但这些工具只是提高了文献管理的效率,并没有从根本上节约研究者的时间和精力,到底使用哪些文献作为引文还是要靠他们自己从海量数据中去选择,而引文推荐(citation recommendation)返回的是一个重要相关文献列表,缩小了研究者的选择范围,在这种背景下,引文推荐研究引起了很多学者的关注。

引文推荐与传统的文献推荐(document recommendation)有着较大的差别。文献推荐是根据用户个性化信息,为用户推送符合其偏好的文献信息[10],引文推荐则是为目标文档或者目标文档中的某个引文上下文(citation context)寻找可供支持的已有研究成果。文档推荐大多是基于用户偏好和文献元数据实现的,而学术文献的引文推荐需要深入到文献内容中,是一种更细粒度单元的文献推荐,其要考虑的特征因素远多于传统的文献推荐。

与传统的学术搜索(academic search)相比,引文推荐对"查询语句"的处理更加多样化、合理化,能够更

* 本文系国家自然科学基金面上项目"面向词汇功能的学术文本语义识别与知识图谱构建"(项目编号: 71473183)研究成果之一。
作者简介: 陈海华(ORCID: 0000-0003-2806-3938),硕士研究生; 孟睿(ORCID: 0000-0003-1580-7627),硕士研究生; 陆伟(ORCID: 0000-0002-0929-7416),副院长,教授,博士生导师,通讯作者,E-mail: weilu@whu.edu.cn。
收稿日期: 2015-06-09 修回日期: 2015-07-21 本文起止页码: 133-143,147 本文责任编辑: 王传清

精确地识别用户需求并进行相关的推荐服务，从而提高研究者的写作效率。近年来，国内外学者就引文推荐进行了一些探索，提出了一些实现算法和模型，并开发了多种引文推荐系统，取得了一定的成果。笔者于 2015 年 5 月 10 日以“citation recommendation”为“Title”或“Abstract”或“Keyword”对 Web of Science、Google Scholar、百度学术进行了检索，以“引文推荐”为“题名”或“摘要”或“关键词”对中国知网、万方、维普、百度学术进行了检索，并利用重要相关文献的参考文献进行二次检索，经过去重、筛选分别得到 49 篇外文文献和 18 篇中文文献（含国内学者发表的外文文献），详细统计见图 1。从图 1 可以看出，目前关于学术文献引文推荐的研究还较少，且国内研究相对滞后。本文希望对已有引文推荐成果进行综述，以期为相关研究提供借鉴和参考。

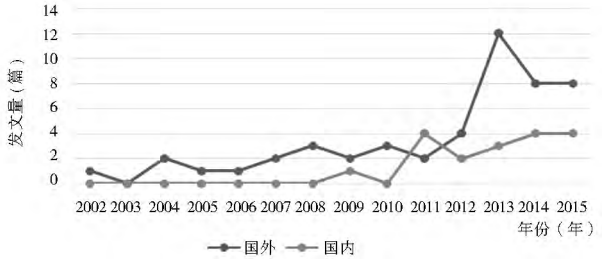

图 1 国内外引文推荐研究发文量年代分布

本文从以下几个方面总结学术文献的引文推荐研究进展：第二节分析引文推荐问题的时间演化过程，第三节从不同角度阐述引文推荐的分类，其中包含两类较为新颖的引文推荐研究，同时介绍一些具有代表性的学术文献引文推荐系统；第四节归纳列举引文推荐研究中常用的 5 种模型与方法；第五节介绍引文推荐的常用数据集和测评方法；最后总结并讨论目前引文推荐研究存在的问题和未来发展方向。

2 引文推荐问题的演化

引文推荐问题起源于 20 世纪初，是一个相对较新的研究方向，在过去 10 年间取得了一定进展，特别是近些年得到学者的广泛关注和重视。为让读者清晰认识引文推荐问题的演化过程，本节分 4 个阶段梳理引文推荐重要成果，分析引文推荐的研究重点和研究方法，见图 2。

自 19 世纪中期第一篇协同过滤算法[11]的文献发表以来，推荐系统成为一个重要的研究领域，很快在商品推荐、电影推荐和新闻推荐上得到了应用，并于 2002 年被 S. M. McNee 等[12]用于传统的文献推荐，且

图 2 引文推荐问题的时间演化

随后的研究大多是基于图模型的。2007 年，T. Strohman 等[13]首次提出引文推荐的概念，认为与信息检索提供给用户短查询相关的文档不同，用户在撰写文献过程中更希望将整篇文献作为查询语句去检索该文献的引文，这是全局引文推荐的雏形，他们结合图模型和文本相似性方法对该问题进行了初步探索，期望引起学者的研究兴趣。

2009 年起，引文推荐研究开始受到关注并进入了稳定发展阶段，引文上下文被应用于引文推荐[14]，使得引文推荐问题的研究深入到引文内容，主题模型[14]、机器学习[15]、翻译模型[16]等方法得以运用，引文推荐的效率大大提高。2009 年，Tang Jie 等[14]首次提出给文中某个具体的引文上下文推荐引文，并应用主题模型完成推荐任务；2010 年，He Qi 等[15]利用引文上下文的差异性首次将引文推荐分为局部引文推荐和全局引文推荐，他认为局部引文上下文是包括引文句在内的若干句子集合，而全局引文上下文是文档的标题和摘要信息，这使得引文推荐研究的问题更加清晰。此前，引文上下文的范围都是人为确定的，2011 年，He Qi 等[17]试图采用机器学习自动识别引文推荐的位置，即在用户不标明具体的引文句时自动识别并给出推荐列表，这种处理方式更加智能化、合理化；然而，Lu Yang 等[16]认为引文上下文与被引文档之间可能存在词汇异质性问题，提出用翻译模型进行解决，从而提高了引文推荐的准确率。

21 世纪 20 年代初，学者们开始重视引文推荐的各个方面，引文推荐研究多点开花。2013 年出现了多样化引文推荐[18]，此时 O. Küçüktunç 等[18-20]意识到之前引文推荐的结果可能过于集中，有必要提供多样化的推荐结果，这更符合用户的实际需求；2013 年，Liu Yaning 等[21]提出个性化引文推荐，在语言模型和翻译模型基础上结合用户偏好，进一步优化引文推荐结果；同年，K. Sugiyama 等[22]对潜在引文推荐进行了探索，2014 年，Tang Xuewei 等[2]针对使用中文撰写文献的作者引用英文文献问题，提出跨语言引文推荐，这两种较

为新颖的引文推荐类型将在第三节进行详细阐述。

2013 年,在 TREC 2013 Knowledge Base Acceleration Track 累积引文推荐(cumulative citation recommendation)子任务的推动下,引文推荐受到了广泛关注,并取得了丰硕成果[23-29]。该任务要求推荐最新研究成果作为引文以协助维基百科词条的更新,这是引文推荐在实践中的具体应用,此前的引文推荐系统包括 Scienstein[30] 和 theadvisor[18],其数据量小、推荐效果差且使用复杂,并未被学者所接受。2014 年,A. Livne 等[1]提出了一个完整的学术文献引文推荐系统框架——CiteSight,综合多种方法进行引文推荐,并运用引文耦合解决了新文献及低被引文献不易被推荐的问题,是未来引文推荐研究及应用的方向。

3 引文推荐的分类

引文推荐的目标是给用户推荐合适的引文,然而用户的需求可能是在文中某处添加需要的引文或获取整篇文档的参考文献列表,也可能是其他一些个性化需求。根据用户的不同需求,本文将引文推荐分成局部引文推荐和全局引文推荐两个主要类别,并介绍跨语言引文推荐和潜在引文推荐两种较为新颖的引文推荐形式。

3.1 局部引文推荐

局部引文推荐(local citation recommendation)是从某个数据集里找出一个有序的文档列表作为文中某个引文上下文的候选引文[15],其目标是为文中某处需要添加引用的引文句推荐引用文献[2],该引文句可以是引文句本身或者由引文句和其前后若干句所组成的引文上下文,局部引文推荐示意图如图 3 所示:

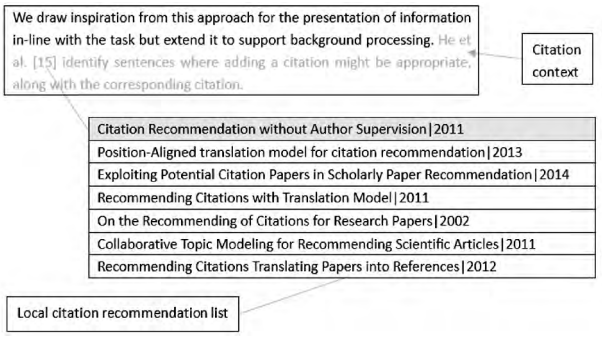

图 3 局部引文推荐示意

引文上下文指的是出现在引证文献正文中引文标识符(citation placeholder)周围、描述被引文献的文字片段[31],其中包含引用标记的句子是显式引文上下文,不包含引文标记的但提供对被引文献进一步描述和补充的句子是隐式引文上下文[32],它们可以提供与引文最直接相关的信息,比如被引文献中具有代表性意义的方法、观点等内容特征。局部引文推荐的关键点是如何确定引文上下文及如何对引文上下文与特定引文进行匹配。S. Bradshaw[33] 用一个固定的窗口——100 个单词(引文前后各 50 个单词)作为引文上下文;P. I. Nakov 等[34] 把引文上下文作为对生物科学文献语义解释的重要工具,他将引文上下文定义为是引文周围的若干句子,但未给出准确的划分范围。A. Ritchie[35] 探讨了引文上下文对于文献检索的影响,实验结果表明使用引文上下文比单纯使用引文句能够提高检索系统的性能。M. A. Angrosh 等[36] 提出基于条件随机场的引文上下文识别,并基于该技术提出了一个提取引文上下文的应用——CitContExt。

在局部引文推荐中,学者们对引文上下文的选择不同,其推荐过程也有差异。He Qi 等[17] 首先提出了一种自动识别文中需添加引文句子的方法,减少了用户手动标记引文的时间,在此基础上实现了一个基于引文上下文的引文推荐系统,他们提出利用概率模型去计算引文上下文和候选引文之间的相似度得分,其不足在于不同的引文上下文可能语义相近但表达完全不同,该模型不能很好地处理此类情况。Tang Jie 等[14] 提出一个基于主题发现的两层 RBM-CS 模型,能够同时发掘引文句和参考文献的主题,并将二者进行匹配。该模型的问题在于,引文句的主题可能不止一个,一处引文可能需要同时引用几篇不同主题的文献,再者该引文句的主题也可能与其他引文句的主题相似,这样就很难准确匹配。另外,某一主题的相关文献数量可能很多,作者并没有提及在这种情况下如何准确筛选出用户所需要的文献。

陆炀[37] 将局部引文推荐问题看成一种信息检索问题,通过在引文上下文和文献之间构建平行语料对的方法训练翻译模型,并将翻译模型整合到经典的语言模型(language model)之中进行检索,试图解决被推荐项与引文上下文(查询语句)之间词汇不匹配的问题。Huang Wenyi 等[38] 将被引文献看成是另一种语言中的一个“新词”,然后利用翻译模型去评估一个引文上下文被“翻译”成引文的概率。刘盛博等[39] 开发了基于引文上下文的引文检索与推荐系统,提取引文信息并建立索引,用户在系统中输入检索词,系统返回相应引文上下文所对应的参考文献,该系统的不足在于只是简单地抽取当前引用句而非完整的引文上下文,且系统只对那些高被引的文献表现出较好的检索性

能。A. Livne 等[1]则引入引文上下文耦合(citation context coupling)的概念来解决那些被引次数较少的文献很难被推荐的问题,他提出基于两步的上下文耦合算法:①首先找到与该文献最相关的若干同被引文献;②找到用来描述以上最相关同被引文献的引文上下文,用 TF - IDF 计算这些引文上下文与目标文献的已有引文上下文之间的相似性,然后进行排名,选取同被引文献排名靠前的引文上下文作为该目标文献的引文上下文,使其引文上下文更加丰富。在此基础之上,设计了能同时支持局部引文推荐和全局引文推荐的推荐系统,用户只要输入文章标题、关键词、主题、摘要等信息并在上下文后面添加引文标记符,系统就会自动地生成局部引文推荐列表和全局参考文献列表,并能动态更新。Zhou Shaoping[2]结合基于内容推荐方法、协同过滤推荐方法和引文分析相关方法实现了一个名为 ActiveCite 的交互式自动引文推荐系统,该系统在局部引文推荐中自动使用当前引文句作为查询语句,在全局引文推荐中自动提取文献主题作为查询语句进行相关引文推荐,其优点在于更加智能化,能尽量减少论文写作过程的中断时间。同时作者从用户角度进行了可用性研究,对引文推荐系统的构建和优化提出了一些建议。

3.2 全局引文推荐

全局引文推荐(global citation recommendation)是指给被推荐文献推荐一个参考文献列表,相比于局部引文推荐,其推荐的范围是整篇文档,涉及的主题较多,但位置上不需要和相关主题一一精确匹配,粒度更粗,在候选集中可检索的范围更广,有些传统的文档推荐方法如随机游走算法和文档相似度计算等可以应用于此。全局引文推荐示意见图 4。

图 4 全局引文推荐示意

含多种信息(内容、作者、引文、协作网络等)的随机游走算法,提出一种面向查询的一元图模型,该模型实现了个性化全局引文推荐。A. Livne 等[1]基于元数据和同被引做全局引文推荐,那些经常同被引的文献被推荐的概率会提高。其他研究[42-43]利用引用偏好和部分参考文献列表来推荐其他的参考文献:R. M. Nallapati 等[42]提出 pairwise-link-LDA 模型和 link-LDA-pLSA 模型来完成该任务,在 Citeseer 数据集上的实验表明其推荐效果比 E. Erosheva 等[44]的基准模型效果更好;S. Bethard 等[43]结合用户的历史引用信息和引用偏好来推荐当前文献的参考文献列表,在 ACL 数据集上准确率为 28.7%,比文本相似性基准线提高了 12.8%。

3.3 其他引文推荐

3.3.1 跨语言引文推荐 传统的引文推荐都是推荐相同语言的相关成果(目前绝大部分成果的文献与对应推荐文献列表均为英文),这种推荐被称为单语言引文推荐(monolingual citation recommendation),而跨语言引文推荐(cross-language citation recommendation)则是为学术文献推荐不同语言的引文。

Tang Xuewei 等[2]对基于上下文感知的跨语言引文推荐进行了研究,为了解决引文上下文与引文之间的语言鸿沟问题,他们提出了双语的引文上下文 - 引文嵌入式算法(简称 BLSRec-I),训练集中包含中文的引文上下文和英文的引文,这样就可以训练出引文上下文和引文最佳匹配位置并嵌入引文。为了提升引文推荐的效果,他们在上述算法中引入翻译结果和摘要信息,分别提出 BLSRec-II 和 BLSRec-III 算法。实验结

果表明,这两种方法在 MAP 和 MRR 上的表现均比基于相似性的方法、上下文相似性模型、翻译模型等基准算法及 BLSRec-I 要好。

跨语言引文推荐本质上与跨语言信息检索(cross-language information retrieval)类似,有些学者也把学术文献的引文推荐看成是信息检索问题[13,45],即把上下文作为查询语句去检索其相关文献。翻译模型经常会被用于跨语言信息检索中,有些模型可能会需要一些另外的资源比如双语词典、机器翻译工具、平行数据集来匹配源语言和目标语言的词语[45-47],这相当于跨语言引文推荐的基础工作。

3.3.2 推荐潜在引文 另一种较为新颖的引文推荐类型就是给学术文献推荐潜在引文(potential citation paper)。研究者们在进行学术创作的过程中所引用的相关文献并不一定是准确的或者完整的,有时某篇文献只涉及与该文献相关但相关度较小的观点或算法等,这在交叉学科中表现得尤为明显,这时使用传统的推荐方法可能会影响系统性能。K. Sugiyama 等[22,48]提出利用学术文献的已有参考文献列表及其内容片段(摘要、引言、结论等)推荐其潜在引文的方法,并对此方法的效果进行了评估,实验表明改进的方法有助于提高文献推荐性能,特别是对于交叉学科中的引文推荐。与以往方法不同,他们更加有效地利用了文献数据:自动识别潜在引文使引文网络不至于过于稀疏;利用协同过滤方法而不仅仅是相关性计算发掘潜在引文。Liu Yaning 等[21]基于用户偏好(用户特征、发表文献记录、引用历史)和内容预测用户引用的潜在文献,实验的准确率相比先前研究在 MAP 和 Recall@10 提高了 27.65% 和 31.67%。A. Livne 等[1]在进行局部引文推荐和全局引文推荐的同时,深入拓展并分析引文的范围,为用户推荐其他未来可能被引用的潜在相关文献。

当然,除了本文介绍的 4 种引文推荐形式,还有其他的引文推荐类型,例如个性化引文推荐、累积引文推荐等,这些都较为常见,在此不复赘述。针对不同的引文推荐类型往往会有不同的推荐方法,有些引文推荐方法也会被应用到各种引文推荐中,例如文本相似度、主题模型等。为了将引文推荐问题阐述得更加清晰,下文重点介绍 5 种引文推荐常用模型与方法。

4 引文推荐的模型与方法

本节总结 5 种引文推荐常用的模型和方法,包括文本相似度(text similarity)、主题模型(topic model)、翻译模型(translation model)、协同过滤算法(collaborative filtering)和混合推荐方法(hybrid recommendation)。

4.1 文本相似度

文本相似度是引文推荐里最常用最基本的方法,使用该方法时,一般是把推荐问题看成是检索问题,即把文本、摘要、引文上下文、用户偏好等作为查询语句,并转化为查询向量去检索相关文档,通过计算查询语句向量和文档向量之间的相关度进行排名。

如 Sun Yueheng 等[49]提出利用基于相似度的算法为学术文献推荐合适的审稿人,他们首先根据审稿人发表文献的信息为审稿人构建偏好向量,将待审文献用向量表示后计算二者的相似度,将相似度较高的文献推荐给相应的审稿人。K. Chandrasekaran 等[50]则用文本相似度的方法为 CiteSeer[51]中的作者推荐相关文献,不同的是,他们用概念树来描述用户偏好,并且通过计算概念树之间的距离来度量用户偏好和文档之间的相似度。

利用文本相似度进行引文推荐的关键在于查询语句的构建。实验表明[13],仅仅使用文本相似度做引文推荐的效果并不理想,因为用户往往会构建一些新词去描述自己的成果,研究同一主题的两个学者也可能对同一概念方法等使用不同的表述,这时使用文本相似度方法就会产生不可靠的结果。另外,基于文本相似度的方法很难充分利用文档质量和权威性等重要特征[13]。

因此,T. Strohman 等[13]在利用文本相似度进行引文推荐时,将文档看作有向图中的一个节点,并充分考虑到作者、引文等信息,他认为这种添加引用信息的文本相似度测量能更真实地反映一个节点的引用情况,同时引入参考文献相似性和 Katz 中心性测量来对候选文档进行排名。A. Livne 等[1]只是将文本相似度用于引文上下文耦合,来丰富文献的引文上下文,进行推荐时则采用了包括机器学习在内的多种方法。He Qi 等[17]结合了语言模型、文本相似性、主题模型、特征依赖模型等去寻找合适的引文上下文。

4.2 主题模型

主题模型常被用于从大量数据中发现潜在主题,该模型提供了一种低维度文档描述。主题模型被广泛应用于语料库挖掘、文本分类和信息检索等领域。

传统的主题模型框架例如 dynamic topic models[52]、pachinko allocation[53]、correlated topic model[54]等忽略了一个重要特征——引用(超链接)[42]。引用不仅反映了两篇文献的主题相似性,同时也反映了被引

文献的权威性，因此将主题模型应用于引文推荐时必须同时考虑以上两个因素。R. M. Nallapati 等[42]介绍了 pairwise-ink-LDA 模型，并提出应用潜在主题模型发掘文本之间的引用关系，这与 D. C. T. Hofmann 等[55]和 E. Erosheva 等[44]提出的模型类似；He Qi 等[17]认为不管是分析文档全文还是将文档看作一条引文，都可以把该文档与主题联系起来，但这两种方式所提炼出的主题可能会不一致；S. Bhatia 等[56]将主题模型用于引文上下文的处理以提升引文推荐的效果。Tang Jie 等[14]进行了基于主题识别的引文推荐研究，他们提出了 RBM-CS(two-layer restricted boltzmann machine)模型来同时发掘学术文献的主题分布和引用关系。在全局引文推荐上，先用主题模型发现文档中的潜在主题，再推荐与这些主题相关的文献；在局部引文推荐上，首先提取引文上下文中的主题，然后给每一个主题推荐最合适的引文，最后将推荐的引文与各个引文句进行匹配。

由于主题模型一般采取迭代算法进行模型训练，往往需要较长的训练时间，所以这种方法不适用于在动态更新的数据集中做引文推荐。

4.3 翻译模型

翻译模型是目前引文推荐中应用最普遍的模型之一。在引文推荐中，往往是将引文上下文和引证文献看成两种不同的“语言”：引文上下文(记作“C”)作为“描述语言”，引证文献(参考文献，记作“D”)作为“参考语言”，训练集记作“T”，然后使用最大似然估计来计算它们之间翻译的概率。

以上只是介绍了一个最简单的基于翻译模型的引文推荐方法，为了提高翻译的准确性，一些学者在此基础上做了相应的改进。2011 年，Lu Yang 等人[43]应用翻译模型进行引文推荐，并比较了该模型用于文章各个章节的效果，实验结果表明，基于摘要的翻译模型比基于全文的翻译模型表现更好，且优于查询似然语言模型和上下文感知相似模型两种基准模型。随后，陆炀等[16,37,57]又提出了加入表征词与词之间联系的翻译模型，用来解决引文与目标文档使用词汇不一致问题(相同语义不同表述)，他们还构建了一个较大的查询语句与相关文档对集合并以此训练出两种翻译模型——全局翻译模型和位置对齐翻译模型。全局翻译模型是在整个训练集上使用查询语句、文档对进行训练，并且分析在文档范围内使用摘要和全文分别会对结果造成怎样的影响；位置对齐翻译模型则是考虑到被引文献不可能通篇谈论的话题都和引文上下文中所谈论的相同，往往只引用被引文献中某处的观点、方法等，这时候就需要一个“位置信息”来把不同的话题隔开[57]。Huang Wenyi 等[58]提出用两种不同语言中对应的词来分别描述引文上下文和引证文献，然后计算该引文上下文映射到引证文献的概率。他们提出平行语料方法，并通过上下文中出现的词找到共引文献，实验证明，他们的方法比此前最佳方法在准确率和召回率上至少提高 5% 和 10%。Tang Yuewei 等[2]也是用翻译模型将中文的上下文与英文的引文匹配，他们还提出结合了机器学习方法的 BLSRec-I、BLSRec-II 和 BLSRec-III 模型，在这种跨语言引文推荐任务中构建翻译模型往往更加复杂。

4.4 协同过滤算法

协同过滤算法的基本思想是根据引用关系或合作者关系等给用户推荐参考文献，一般用于全局引文推荐。由于传统的协同过滤推荐算法几乎不会推荐被引较少的文献，因此在推荐较新文献上具有较为明显的局限性[58]，学者们在使用该方法进行引文推荐时都会有所改进。Ding Zhou 等[59]提出一种半监督学习方法并结合多种关系图(包括引用、作者、位置等信息)来评测文档相似性；Wang Chong 等[60]提出一个协同过滤主题回归模型(collaborative topic regression，简称“CTR”)进行学术文献推荐，该模型结合了传统协同过滤和概率主题模型的优点；K. Sugiyama 等[22,48]为了解决新发表文献在引文网络中权重较低及交叉学科中引文推荐效果较差的问题，运用协同过滤方法识别潜在引文，他们还提出基于缺失值的协同过滤框架。S. M. McNee 等[12]利用文献之间的引文网络、文献引用信息、共引信息构建评分矩阵进行协同引文推荐，他们认为经常同被引的文献更容易被推荐。C. Basu 等[61]则采用审稿人多轮打分的方式，并且用扩展贝叶斯和 K 邻近算法进行协同过滤文献推荐，实验发现利用的信息越多推荐效果越好。

然而，协同过滤算法并未深入到引文内容，近年来较少被单独用于引文推荐，往往会与其他方法一起被使用。

4.5 混合推荐方法

不同的模型和推荐方法都有其自身的不足，有些算法在单独使用的情况下并不能完成整个引文推荐任务，需要结合其他方法，才能达到更好的推荐效果。因此，在实际引文推荐中，经常会使用混合推荐方法，融合主题模型、翻译模型、机器学习、协同过滤和其他一些较为有效的方法以得到最终推荐结果。

Huang Zan 等[62]提出基于一元图模型的文献推荐系统,该模型不仅表达了文献之间的关联和用户之间的关联,同时两层关联之间靠用户对文献打分进行连接,图搜索技术被应用到该系统中。G. Bela 等[30]介绍了世界上第一个结合基于内容和基于协同过滤两种方法的混合推荐系统 Scienstein,该系统集成了引文分析、作者分析、源分析、隐式打分、显式打分等多方面信息。R. Torres 等[63]介绍了一种结合协同过滤和基于内容过滤的混合推荐算法,实验表明混合推荐算法要优于单独的推荐算法。M. Gori 等[40]提出了一个基于学术引用图和随机游走的学术文献推荐算法;Zhang Ming 等[64]实现了先用协同过滤发现相似邻居,再用基于内容推荐技术进行文献推荐的混合推荐系统;石杰[65-66]等提出基于多因素的引文推荐策略,此处的多因素包括引文自身因素和用户因素,其本质仍然是用户协同过滤和相似度计算相结合的混合推荐。A. Livne[1]在引文推荐的不同步骤使用不同方法,充分利用文本相似度、机器学习、协同过滤等方法的优点解决了低被引文献难被推荐和推荐结果过于集中等问题,同时提升了引文推荐的速率,满足了不同水平用户的需求,值得学习和借鉴,这种混合引文推荐方法是未来引文推荐研究的方向。

以上详细介绍了当前学术文献引文推荐中5种常用的推荐模型与方法,总的来说,各有优劣。文本相似度、主题模型和翻译模型均属于基于内容的推荐方法,其局限性在于推荐的结果可能过于专门化或重复冗余,且不能区分内容上相近的重要文献与一般文献;协同过滤方法由于可利用的数据远少于待推荐的数据,一般会遇到稀疏的问题,与基于内容的推荐方法一样,在新用户或低被引文献上表现不好;而混合推荐方法整合了二者的不足,充分利用了引文内容、用户偏好、引文网络等信息,在 CiteSight、RefSeer 等引文推荐系统中表现出了很好的效果。随着引文推荐问题研究的深入,应结合多种方法以保证推荐引文的权威性、多样性、新颖性,同时提高引文推荐的准确性与速率。

5 引文推荐数据与评价

引文推荐本质上仍是信息检索的范畴,研究者往往要根据不同的推荐任务在具体数据集上进行实验,给出相应的引文推荐结果,并对结果进行评测,才能说明其推荐思路及方法的可行性与高效性。笔者将从引文推荐数据和引文推荐评价两方面进行梳理和总结。

5.1 引文推荐数据

目前引文推荐研究的角度和方法十分多样,所以引文推荐的数据来源也多种多样,但仍有部分数据集得到学者的广泛使用。目前在引文推荐中,常用的数据集包括 CiteSeer(CiteSeer X)、CiteULike、NIPS 等。CiteSeer 数据集被 He Qi 等[15,17]、K. Chandrasekaran 等[50]、S. Bhatia 等[56]、Tang Jie 等[14]、R. M. Nallapati 等[42]、B. Wellner 等[67]用于引文推荐,例如,Tang Jie 等[14]选取了 CiteSeer 网站中包含 32 558 条引用关系的 3 335 篇文章作为数据集,其中不包含没有出现在该数据集中的参考文献,经过对停用词、数字和出现次数较少词的过滤最终得到了 634 875 个词用于基于主题的引文推荐。CiteSeer 数据集被广泛应用的原因是:①能够解析出文章的引文并且识别表述不同但语义相同的引文,确保得到一篇文献的所有引文;②能够提取文章的引文上下文,使研究者了解其他学者对该文章的观点;③能够通过公共引文和文档相似性识别相关文献。CiteULike 则被 S. Bhatia 等[56]、Tang Jie 等[14]用于引文推荐,他们使用的数据包括文档数、引文上下文数量、引文上下文的非重复单词数量、参考文献、平均引文数量等信息。

另外一些数据在引文推荐中也得到很好的应用。TREC2013 累积引文推荐子任务使用 KBA Stream Corpus 2012[68]作为数据,该数据集包括 2011 年 10 月到 2014 年 4 月的新闻、微博、论坛和 bitly.com 上的链接数据等 3 种文档类型,目的是协助维基百科工作人员完成词条的更新。其他一些成果[1-2,13-14,16,69]根据特定的研究任务(如跨语言引文推荐、上下文耦合引文推荐、预测引文数量等)选择了诸如 TREC、ACM 等数据集或计算机科学、自然语言处理、信息检索与数据挖掘等领域的核心期刊或谷歌学术、微软学术等数据库中若干经过相应预处理的文档集合作为数据集。

目前引文推荐数据集存在的问题在于规模较小,且局限在某一领域。相比来说,综合的大数据集噪音更多,处理起来更加困难,且对系统的要求较高,这也是将学术文献引文推荐推向应用的一大瓶颈。

5.2 引文推荐评价

引文推荐的任务不同,其评价标准也不统一,但大体上一般从以下 3 个方面进行评测:①对单个推荐

的评测;②对多个推荐的评测(通常用于对推荐系统的评价);③面向用户的评测。目前研究者大多只关心前两者的评测,而缺乏面向用户的评测标准。

对引文推荐进行评测的指标也是采用信息检索评价的常用指标,主要包括:Recall[2,15,60,70]、Precision[14,38]、F-measure[23-29,38]、Bpref(binary preference measure)[14,38]、MRR(mean reciprocal rank)[2,14,48]、MAP(mean average precision)[2,14,16,21,37]、NDCG(normalized discounted cumulative gain)[1,11,17-20,22,48]等,这些指标在TREC、SIGIR、NTCIR、iConference、ACL等国际知名会议中也被广泛使用。下面对这些指标在引文推荐中的计算方法进行详细介绍。

在一个测试集T中,假设原始的引文列表为Rg,而实际推荐的引文列表为Rr,则正确的推荐结果为$R_g \cap R_r$,此时Recall、Precision、F-measure的计算公式如下[38]:

$$\text{Recall} = \frac{|R_g \cap R_r|}{R_g}, \text{Precision} = \frac{|R_g \cap R_r|}{R_r},$$

$$F = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

然而,引文推荐系统的效率也与推荐结果的顺序有关,例如更相关的文献应该排在前面,这在Recall、Precision、F-measure中并不能反映出来。为解决该问题,引入Bpref和MRR两个指标。假设对于一个引文上下文或一篇文献某推荐算法给出的推荐列表为S,其中正确的推荐结果集合为R,若r为S中一条正确的推荐结果,i为S中一条错误的推荐结果,rankq为S中第一条正确结果的排名,则Bpref和MRR的计算公式如下[38]:

$$\text{Bpref} = \frac{1}{|R|} \sum_{r \in R} 1 - \frac{|i \text{ ranked higher than } r|}{|S|}$$

$$\text{MRR} = \frac{1}{|R|} \sum_{q \in T} 1 - \frac{1}{\text{rank}_q} \quad (2)$$

在局部引文推荐中,某个引文上下文可能会有多篇引文;全局引文推荐中,文档级别的参考文献也不止一篇,MAP考虑的是每篇被推荐文档准确率的平均值,假设R(di)表示推荐列表中第i篇文档di是否相关(1表示相关,0表示不相关),则MAP的计算公式为[16]:

$$\text{MAP}(d_1, d_2, \cdots, d_N) = \frac{\sum_i \frac{R(d_i)}{i} \sum_{j \leq i} R(d_i)}{\sum_i R(d_i)} \quad (3)$$

事实上,推荐的每条引文不仅仅只有相关和不相关两种情况,而是有相关度级别的,比如0、1、2、3等。一般认为对于给出推荐列表:相关度级别越高的结果越多越好,且越靠前越好。NDCG正是考虑到相关度级别,故在引文推荐评测中得到了广泛应用,推荐列表中位置i的NDCG值计算公式为[17]:

$$\text{NDCG@}i = Z_i \sum_{j=1}^{i} \frac{2^{r(j)} - 1}{\log(1 + j)} \quad (4)$$

其中Z_i为归一化常数,r(j)为推荐列表中第j篇文档的等级得分。

以上的评测指标都没有考虑用户因素,在实际应用中,推荐结果的相关与不相关是由用户判定的。因此,未来引文推荐的评测应将用户评价指标考虑在内,从而构建一套科学完整的学术文献引文推荐评价体系。

6 总结与展望

6.1 对研究现状的总结

学术文献的引文推荐力求使研究者花最少的精力找到最适合的参考文献和引文,在减轻用户寻找文献负担的同时避免遗漏重要相关参考文献。本文从引文推荐的分类、引文推荐的方法、引文推荐的数据集和测评方法等几个方面介绍了学术文献引文推荐的最新研究进展。虽然目前针对学术引文推荐这一研究产生了丰硕的成果,但仍存在着一些问题:

(1)没有考虑用户偏好的动态变化性及研究领域的综合性。研究热点在不断更新,用户兴趣也在不断变化;再者,随着交叉学科的日益增多,用户的研究兴趣可能涉及多个领域,在论文中反映的主题也可能呈现多元化,如何构建动态的、合理的用户偏好模型值得思考。

(2)目前,大部分引文推荐系统尚处于实验阶段,还没有被推向实际应用。一些系统可能在小数据集上表现出很好的推荐效果,但缺乏在大数据或多领域环境下的系统研究。同时关于引文推荐系统的相关用户研究也十分匮乏。

6.2 对未来学术文献引文推荐研究方法的思考

通过对研究现状的总结和思考,笔者认为未来关于学术文献引文推荐应该根据具体的推荐任务综合运用各种方法,尽可能发挥其优点,提高推荐的质量与效率,例如在全局引文推荐中可以结合协同过滤、用户偏好模型和主题模型,并给出多样化的推荐列表,避免推荐结果重复冗余或过于单一;在局部引文推荐中应结合具体的引文上下文,并考虑到引文

上下文与被引文献之间的词汇异质性综合应用翻译模型、主题模型等方法,在引文上下文自动识别时可结合文本相似性、上下文耦合、机器学习等方法和技术。另外,引文推荐评测过程中不仅要进行信息检索实验的相关评价,还要进行相关的用户研究,这样才能更快地将引文推荐推向实际应用。

6.3 未来学术文献引文推荐的潜在研究方向

6.3.1 引文更加有效的语义化表达方法 引文上下文增强了被引文献的表达效果,改善了引文推荐的效果。但被引频次较低的文献很难单纯通过引文上下文得到很好的描述,往往要借助于引文之间的耦合关系,这时就很有必要探索如引文上下文耦合等方法改进被推荐文献的描述。

6.3.2 引文上下文(查询语句)与被推荐文献之间词汇的异质性(不匹配)问题 在信息检索和推荐系统中,用户构建查询语句时使用的词汇多种多样,往往与参考文献不一致,目前,有学者试图从文本集合中挖掘词汇之间的关系来解决该问题[71],然而他们大多只关注查询扩展能提高检全率却很少关注该方法带来的推荐主题偏移问题,会导致推荐不相关的引文。尽管陆炀[37]在基于翻译模型的引文推荐中提出用全局翻译模型和位置对齐翻译模型解决该问题,取得了很好的效果,但这方面的研究还有待丰富和深入。

6.3.3 结合自然语言生成技术的引文推荐 此前,很多学者进行了文本自动摘要的相关研究[72-74],取得了丰硕成果。如何将该技术用于学术文献引文推荐,使推荐结果不仅仅是文献列表,还包括推荐文献的简要摘要内容,是研究者应该考虑的问题。

6.3.4 跨语言引文推荐研究亟待丰富 目前只有Tang Xuewei等[2]对跨语言引文推荐研究进行了初步探索,且目前成果仅仅是为中文文献推荐英文引文,其他语言的成果推荐仍不支持。

参考文献:

[1] Livne A, Gokuladas V, Teevan J, et al. CiteSight: Supporting contextual citation recommendation using differential search [EB/OL]. [2015-04-25]. http://dl.acm.org/citation.cfm?id=2609585.

[2] Tang Xuewei, Wan Xiaojun, Zhang Xun. Cross-language context-aware citation recommendation in scientific articles [EB/OL]. [2015-04-25]. http://dl.acm.org/citation.cfm?id=2609564.

[3] Wouters P. The citation culture [D]. Amsterdam: University of Amsterdam, 1999.

[4] May R M. The scientific wealth of nations [J]. Science, 1997, 275(5301): 793-796.

[5] Mallik A, Mandal N. Bibliometric analysis of global publication output and collaboration structure study in microRNA research [J]. Scientometrics, 2014, 98(3): 2011-2037.

[6] LaTeX. A document preparation system [EB/OL]. [2015-04-25]. http://www.latex-project.org/.

[7] Liu Bin, Guo S D. Attentions on Grey System Theories by China scholars——Based on literature metrology during 1982-2009 [J]. Journal of Grey System, 2010, 22(2): 137-146.

[8] Zhao F H, Yang B. Improving the efficiency of sci-tech novelty search based on EndNote [C]//10th International Conference on Innovation and Management. Wuhan: Wuhan University of Technology Press, 2013: 559-562.

[9] Zhou Shaoping. ActiveCite: An interactive system for automatic citation suggestion [D]. Singapore: National University of Singapore, 2010.

[10] 李琳娜,张志平,刘春霞. 文献推荐系统综述[J]. 数字图书馆论坛,2012(5): 32-37.

[11] Resnick P, Iacovou N, Suchak M, et al. GroupLens: An open architecture for collaborative filtering of netnews [EB/OL]. [2015-04-25]. http://dl.acm.org/citation.cfm?id=192905.

[12] McNee S M, Albert I, Cosley D, et al. On the recommending of citations for research papers [EB/OL]. [2015-04-25]. http://dl.acm.org/citation.cfm?id=587096.

[13] Strohman T, Croft W B, Jensen D. Recommending citations for academic papers [EB/OL]. [2015-04-25]. http://dl.acm.org/citation.cfm?id=1277868.

[14] Tang Jie, Zhang Jing. A discriminative approach to topic-based citation recommendation [M]//Advances in Knowledge Discovery and Data Mining. Berlin: Springer Berlin Heidelberg, 2009: 572-579.

[15] He Qi, Pei Jian, Kifer D, et al. Context-aware citation recommendation [EB/OL]. [2015-04-25]. http://dl.acm.org/citation.cfm?id=1772734.

[16] Lu Yang, He Jing, Shan Dongdong, et al. Recommending citations with translation model [EB/OL]. [2015-04-25]. http://dl.acm.org/citation.cfm?id=2063879.

[17] He Qi, Kifer D, Pei Jian, et al. Citation recommendation without author supervision [EB/OL]. [2015-04-25]. http://dl.acm.org/citation.cfm?id=1935926.

[18] Küçüktunç O, Saule E, Kaya K, et al. Result diversification in automatic citation recommendation [EB/OL]. [2015-04-25]. https://scholar.google.com/scholar?q=Result+diversification+in+automatic+citation+recommendation&btnG=&hl=zh-CN&as_sdt=0%2C5.

[19] Küçüktunç O, Saule E, Kaya K, et al. Diversified recommendation on graphs: Pitfalls, measures, and algorithms [EB/OL].

[2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 2488451.

[20] Küçüktunç O, Saule E, Kaya K, et al. Diversifying citation recommendations [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2014, 5(4): 1 - 19.

[21] Liu Yaning, Yan Rui, Yan Hongfei. Guess what you will cite: Personalized citation recommendation based on users' preference [EB/OL]. [2015 - 04 - 25]. http://link.springer.com/chapter/10.1007/978 - 3 - 642 - 45068 - 6_37.

[22] Sugiyama K, Kan M Y. Exploiting potential citation papers in scholarly paper recommendation [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 2467701.

[23] Gebremeskel G G, He Jie, De Vries A P, et al. Cumulative citation recommendation: A feature-aware comparison of approaches [EB/OL]. [2015 - 04 - 25]. http://ieeexplore.ieee.org/xpls/abs_all.jsp? arnumber = 6974848&tag = 1.

[24] Balog K, Ramampiaro H, Takhirov N, et al. Multi-step classification approaches to cumulative citation recommendation [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 2491775.

[25] Balog K, Ramampiaro H. Cumulative citation recommendation: Classification vs. ranking [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 2484151.

[26] Dietz L, Dalton J, Balog K. Time-aware evaluation of cumulative citation recommendation systems [EB/OL]. [2015 - 04 - 25]. https://scholar.google.com/scholar? q = Time - aware + evaluation + of + cumulative + citation + recommendation + systems&btnG = &hl = zh - CN&as_sdt = 0%2C5.

[27] Abbes R, Pinel-Sauvagnat K, Hernandez N, et al. IRIT at TREC knowledge base acceleration 2013: Cumulative citation recommendation task [EB/OL]. [2015 - 04 - 25]. https://hal.archives - ouvertes.fr/hal - 01143717/.

[28] Wang Jingang, Song Dandan, Wang Qifan, et al. An entity class-dependent discriminative mixture model for cumulative citation recommendation [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 2767698.

[29] Wang Jingang, Liao Lejian, Song Dandan, et al. Resorting relevance evidences to cumulative citation recommendation for knowledge base acceleration [M]//Web-Age Information Management. Berlin: Springer International Publishing, 2015: 169 - 180.

[30] Bela G, Jöran B, Christian H. Scienstein: A research paper recommender system [C]//Proceedings of the International Conferenceon Emerging Trends in Computing. Virudhunagar: Kamaraj College of Engineering and Technology India, 2009: 309 - 315.

[31] 刘洋，崔雷. 引文上下文在文献内容分析中的信息价值研究[J]. 图书情报工作，2014，58(6)：101 - 104.

[32] 雷声伟. 学术文献的引文上下文探测研究[D]. 武汉：武汉大学，2015.

[33] Bradshaw, S. Reference directed indexing: Redeeming relevance for subject search in citation indexes [C]//Proceedings of the 7th European Conference, ECDL'03. Berlin: Springer Berlin Heidelberg, 2003: 499 - 510.

[34] Nakov P I, Schwartz A S, Hearst M. Citances: Citation sentences for semantic analysis of bioscience text [EB/OL]. [2015 - 04 - 25]. https://scholar.google.com/scholar? q = Citances%3A + Citation + sentences + for + semantic + analysis + of + bioscience + text&btnG = &hl = zh - CN&as_sdt = 0%2C5.

[35] Ritchie A. Citation context analysis for information retrieval [D]. Cambridge: University of Cambridge, 2008.

[36] Angrosh M A, Cranefield S, Stanger N. Conditional random field based sentence context identification: Enhancing citation services for the research community [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 2527216.

[37] 陆炀. 基于翻译模型的引文推荐[D]. 北京：北京大学，2013.

[38] Huang Wenyi, Saurabh K, Cornelia C, et al. Recommending citations: Translating papers into references [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 2398542.

[39] 刘盛博，丁堃，刘则渊. 基于引用内容的引文检索与推荐系统[J]. 情报学报，2013，32(11)：1157 - 1163.

[40] Gori M, Pucci A. Research paper recommender systems: A random-walk based approach [EB/OL]. [2015 - 04 - 25]. http://ieeexplore.ieee.org/xpls/abs_all.jsp? arnumber = 4061472.

[41] Meng Fanqi, Gao Dehong, Li Wenjie, et al. A unified graph model for personalized query-oriented reference paper recommendation [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 2507831.

[42] Nallapati R M, Ahmed A, Xing E P, et al. Joint latent topic models for text and citations [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 1401957.

[43] Bethard S, Jurafsky D. Who should I cite: Learning literature search models from citation behavior [EB/OL]. [2014 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 1871517.

[44] Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications [J]. Proceedings of the National Academy of Sciences, 2004, 101(suppl 1): 5220 - 5227.

[45] Nie J Y, Simard M, Isabelle P, et al. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 312656.

[46] Grefenstette G. The problem of cross-language information retrieval [M]//Cross-Language Information Retrieval. Berlin: Springer US, 1998: 1 - 9.

[47] Gollins T, Sanderson M. Improving cross language retrieval with triangulated translation [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm? id = 383965.

[48] Sugiyama K, Kan M Y. A comprehensive evaluation of scholarly

paper recommendation using potential citation papers [J]. International Journal on Digital Libraries, 2014, 16(2): 91 - 109.

[49] Sun Yueheng, Ni Weijie, Men Rui. A personalized paper recommendation approach based on web paper mining and reviewer's interest modeling [EB/OL]. [2015 - 04 - 25]. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5401291.

[50] Chandrasekaran K, Gauch S, Lakkaraju P, et al. Concept-based document recommendations for citeseer authors [C]//Adaptive Hypermedia and Adaptive Web - Based Systems. Berlin: Springer Berlin Heidelberg, 2008: 83 - 92.

[51] Giles C L, Bollacker K D, Lawrence S. CiteSeer: An automatic citation indexing system [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=276685.

[52] Blei D M, Lafferty J D. Dynamic topic models [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=1143859.

[53] Li Wei, McCallum A. Pachinko allocation: DAG-structured mixture models of topic correlations [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=1143917.

[54] Lafferty J D, Blei D M. Correlated topic models [EB/OL]. [2015 - 04 - 25]. https://scholar.google.com/scholar?q=Correlated+topic+models&btnG=&hl=zh-CN&as_sdt=0%2C5.

[55] Hofmann D C T. The missing link-A probabilistic model of document content and hypertext connectivity [C]//Proceedings of the 2000 Conference on Advances in Neural Information Processing Systems. Massachusetts: The MIT Press, 2001: 430 - 436.

[56] Bhatia S, Kataria S, Mitra P. Utilizing context in generative bayesian models for linked corpus [EB/OL]. [2015 - 04 - 25]. http://www.aaai.org/ocs/index.php/AAAI/AAAI10/index.

[57] He Jing, Nie Jian Yun, Lu Yang, et al. Position-aligned translation model for citation recommendation [C]//String Processing and Information Retrieval. Berlin: Springer Berlin Heidelberg, 2012: 251 - 263.

[58] 丁彬钊. 基于引文信息的协同过滤算法研究 [D]. 长春: 吉林大学, 2011.

[59] Ding Zhou, Zhu Shenghuo, Yu Kai, et al. Learning multiple graphs for document recommendations [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=1367517.

[60] Wang Chong, Blei D M. Collaborative topic modeling for recommending scientific articles [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=2020480.

[61] Basu C, Hirsh H, Cohen W W, et al. Technical paper recommendation: A study in combining multiple information sources [J]. Journal of Artificial Intelligence Research, 2002, 14: 241 - 262.

[62] Huang Zan, Chung Wingyan, Ong T H, et al. A graph-based recommender system for digital library [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=544231.

[63] Torres R, McNee S M, Abel M, et al. Enhancing digital libraries with TechLens+ [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=996402.

[64] Zhang Ming, Wang Weichun, Li Xiaoming. A paper recommender for scientific literatures based on semantic concept similarity [M]//Digital Libraries: Universal and Ubiquitous Access to Information. Berlin: Springer Berlin Heidelberg, 2008: 359 - 362.

[65] 石杰, 申德荣, 聂铁铮, 等. 一种基于多因素的引文推荐方法 [J]. 计算机研究与发展, 2011 (S3): 180 - 188.

[66] 石杰. 基于多因素的引文推荐策略研究 [D]. 沈阳: 东北大学, 2011.

[67] Wellner B, McCallum A, Peng F, et al. An integrated, conditional model of information extraction and coreference with application to citation matching [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=1036915.

[68] KBA stream corpus 2012 [EB/OL]. [2015 - 4 - 10]. http://trec-kba.org/kba-stream-corpus-2012.shtml.

[69] Yan Rui, Tang Jie, Liu Xiaobing, et al. Citation count prediction: Learning to estimate future citations for literature [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=2063757.

[70] Shaparenko B, Joachims T. Identifying the original contribution of a document via language modeling [M]//Machine Learning and Knowledge Discovery in Databases. Berlin: Springer Berlin Heidelberg, 2009: 350 - 365.

[71] Karimzadehgan M, Zhai Cheng Xiang. Estimation of statistical translation models based on mutual information for ad hoc information retrieval [EB/OL]. [2015 - 04 - 25]. http://dl.acm.org/citation.cfm?id=1835505.

[72] Mani I, Maybury M T. Advances in automatic text summarization [J]. Compare, 1998, 32(3): 88 - 90.

[73] Das D, Martins A F T. A survey on automatic text summarization [J]. Literature Survey for the Language and Statistics II course at CMU, 2007, 4: 192 - 195.

[74] Kim S N, Medelyan O, Kan M Y, et al. Automatic keyphrase extraction from scientific articles [J]. Language resources and evaluation, 2013, 47(3): 723 - 742.

作者贡献说明:

陈海华: 参与研究思路修改讨论和研究框架整理,进行文献调研,撰写和修改论文;

孟睿: 参与讨论修改研究思路和研究框架,修改论文并校对;

陆伟: 提出研究思路和研究整体框架,参与论文修改。

(下转第147页)

一种必然，也是唯一出路！否则，失去生存发展价值、被边缘化、被抛弃，绝不是危言耸听！本次年会以“转型”为焦点，希望图书情报机构能认真审视自身的优势和特点、自身的劣势和不足，多样性、差异化发展，在浩瀚的科技信息服务市场中，寻找好自身的定位，谋划好自身的发展路径。年会得到了中国图书进出口（集团）总公司、汤森路透知识产权与科技集团（中国区）的鼎力支持。《现代图书情报技术》、《图书情报工作》、《中国文献情报》（英文刊）提供了媒体支持。

赵树宜
（中国图书馆学会专业图书馆分会秘书处）
揭玉斌
（中国化工信息中心）

（上接第 143 页）

Research Review on Citation Recommendation of Academic Literatures

Chen Haihua　Meng Rui　Lu Wei
School of Information Management, Wuhan University, Wuhan 430072
Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan 430072

Abstract: [Purpose/significance] Citation Recommendation of academic literature refers to automatically recommend suitable citations and references for a given paper. With the aid of citation recommendation, authors can improve the efficiency of writing academic literatures and avoid missing important relevant literatures. [Method/process] This paper analyzesthe latest development of researches on citation recommendation at home and abroad, describes the evolution of citation recommendation, and elaborates the citation recommendationfromdifferent perspectives, such asthe local citation recommendation and global citation recommendation. Main citation recommendation techniques including text similarity, topic model, translation model, collaborative filtering and hybrid recommendation are reviewed. Also, several major data – sets and evolution methods on citation recommendation are summarized. [Result/conclusion] Previous researches didn't consider the dynamic tendency of user's preference and the comprehensiveness of research area. What's more, user studies and practical applications of citation recommendation are still lacking. To fill the gap, future works on citation recommendation may focus on semantic expression, natural language generation and cross – language citation recommendation.

Keywords: citation recommendation　citation recommendation classification　citation recommendation method　citation context