面向搜索引擎的实体推荐综述

黄际洲[1),2)] 孙雅铭[2)] 王海峰[2)] 刘 挺[1)]

[1)](哈尔滨工业大学计算机学院社会计算与信息检索研究中心 哈尔滨 150001)

[2)](百度公司 北京 100085)

摘 要 面向搜索引擎的实体推荐任务旨在为用户输入的搜索查询推荐出相关实体,从而帮助用户发现感兴趣的实体,提升用户的搜索体验.此外,为了帮助用户更好地理解实体推荐结果,还需要为被推荐的实体集合以及每一个被推荐实体生成恰当且合理的推荐理由.实体推荐能够帮助用户便捷地获得与其搜索需求相关的信息,有助于提升用户的信息发现体验,因此已成为现代搜索引擎中必不可少的功能之一.与传统领域的推荐任务相比较,面向搜索引擎的实体推荐面临更多的挑战,例如搜索查询中实体指称的歧义性以及实体推荐的领域无关性等.针对搜索引擎实体推荐任务的特点与存在的挑战,我们认为构建一个完备的实体推荐系统需要解决如下三个子研究任务:实体链接、实体推荐与推荐理由生成.实体链接任务的目标是将搜索查询中的实体指称消除歧义并链接到知识库中无歧义的实体上,以获得与搜索查询对应的查询实体.实体推荐任务的目标是获取与查询实体相关的实体集合并对其进行排序.为了提供更准确的推荐结果,往往还需要进一步利用历史搜索信息获取用户对实体的偏好并对当前查询进行更好地理解.推荐理由生成任务的目标是为被推荐的实体集合以及每一个被推荐实体生成推荐理由,其中集合推荐理由解释的是该集合中的被推荐实体与查询实体的关系,实体推荐理由则是单个实体被推荐的理由.本文首先介绍面向搜索引擎的实体推荐任务的研究背景与意义、存在的挑战以及各子任务,然后详细介绍每一个子任务存在的技术挑战、研究现状以及解决方法,最后对未来研究方向进行展望并对本文进行总结.

关键词 搜索引擎;实体推荐;实体链接;推荐理由

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.01467

A Survey of Entity Recommendation in Web Search

HUANG Ji-Zhou[1),2)] SUN Ya-Ming[2)] WANG Hai-Feng[2)] LIU Ting[1)]

[1)](*Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin* 150001)

[2)](*Baidu Inc., Beijing* 100085)

Abstract Entity recommendation aims to provide search users with entity suggestions relevant to their information needs, which can help them to explore and discover entities of interest. For this reason, over the past few years, major commercial Web search engines have proactively recommended related entities for a query along with the regular Web search results to enrich and improve the user experience of information retrieval and discovery. To help users better understand why the entities are recommended to them, it is also important to provide explanations for recommendations. The task of building an entity recommendation system presents more challenges than the task of building a traditional item-based recommender system because of the ambiguity of the entities mentioned in queries, the domain-agnostic recommendation methods for Web-scale queries, and the cross-domain recommendation scenarios. To address these challenges, the following three sub-tasks should be studied on building an entity recommendation system in Web

收稿日期:2018-11-22;在线出版日期:2019-03-22. 本课题得到国家"九七三"重点基础研究发展计划项目基金(2014CB340505)资助.
黄际洲,博士研究生,主要研究方向为自然语言处理、推荐系统、人工智能. E-mail: huangjizhou01@baidu.com. **孙雅铭**,博士,工程师,主要研究方向为实体消歧、自然语言处理. **王海峰**,博士,教授级高工,博士生导师,主要研究领域为自然语言处理、机器翻译、人工智能. **刘 挺**,博士,教授,博士生导师,主要研究领域为人工智能、自然语言处理、社会计算.

search engines. The first is entity linking in queries, which aims to disambiguate the entity mentioned in a query and link it to the corresponding entity in a knowledge base. To improve the entity linking accuracy, an entity linking system should consider additional information such as the query context and a user's search history. The second is entity recommendation, which aims to find a set of related entities to a query, and then rank these entities. Specifically, an entity recommendation model typically consists of two components: related entity finding and entity ranking. The former extracts a set of candidate entities related to a query that a user is searching for, while the latter ranks the candidate entities according to how well they meet the user's information need. To better understand a user's information needs and capture a user's preferences, an entity recommendation model should exploit additional information such as a user's search history. There are two kinds of search history: short-term search history in a single session and long-term search history across all sessions. The short-term search history, which consists of in-session preceding queries and clickthrough data, can be exploited to help understand a user's information needs and capture a user's interests on entity preference in the current session. The long-term search history includes query history and clickthrough data across all sessions for a period of time, which reflects a user's interests accumulated over time and could be used to capture the user's intrinsic interests on entity preference. Therefore, in order to generate more relevant entity recommendations w. r. t. the user's information needs and preferences, it is important for an entity recommendation model to exploit as many search histories as possible. The third is recommendation captioning, which aims to explain why two entities are related and why a group of entities is recommended to a user. Presenting related entities with plausible explanations can help users quickly figure out the connections between the query and the recommended entities as well as the key facts of these entities, which in turn increases the understandability of the recommendations and user engagement. In this paper, the research background and the challenges of this task are presented first, and then the related studies and methods are introduced. Finally, problems are discussed, and several future research directions are suggested.

Keywords search engine; entity recommendation; entity linking; recommendation captioning

1 引 言

搜索引擎是用户获取信息的重要工具.近年来,为了更好地满足用户的信息获取需求,搜索引擎从最初只能被动地根据用户输入的搜索查询(search query)返回相关网页,逐步改进到能够主动为用户提供直接答案[1-2]和推荐相关信息[3-6].用户对实体的信息需求较大,例如超过70%的搜索查询包含命名实体(named entity)[7],在所有搜索查询中大约40%的搜索查询其主要搜索需求为其中出现的一个实体[8].大规模知识库(knowledge base)如Freebase[9]、DBpedia[10]的出现使得搜索引擎可以为搜索查询中的核心实体推荐相关的实体.面向搜索引擎的实体推荐(为简便起见,后续统一简称为实体推荐)不仅能够帮助用户探索和发现感兴趣的相关实体,而且对于提升用户参与度(user engagement)具有至关重要的作用.实体推荐已经成为现代搜索引擎必不可少的功能之一.

实体推荐系统的目标是根据用户输入的查询,在搜索结果中提供相关实体建议,以帮助用户发现更多与其搜索需求相关的信息.图1显示了百度搜索引擎为查询"奥巴马"所提供的搜索结果.在搜索结果页中,左侧区域展现的是与该查询相关的网页,而与该查询相关的实体推荐,则展现在右侧区域的"相关人物"中,每一个被推荐的实体还附有一条恰当且合理的推荐理由以便让用户迅速了解被推荐的实体.这些由系统推荐的实体,能够帮助用户便捷地找到与其搜索需求相关的其他实体,让用户多了一种探索更多信息的方式,能够有效提升用户的信息发现体验.

图 1　查询“奥巴马”所对应的百度搜索结果

与传统推荐任务相比，面向搜索引擎的实体推荐任务主要存在以下挑战：(1) 在传统推荐任务中，用户感兴趣的物品(item)是显式的和具体的，例如某一个商品或电影，而在搜索引擎的实体推荐任务中，用户所感兴趣的实体并没有被显式地给出. 查询中的实体指称(mention)通常具有歧义，例如“奥巴马”在百度百科中有 7 个不同的义项，因此需要对实体指称进行消歧以获取用户的搜索需求；(2) 传统推荐任务中候选推荐物品的规模远小于搜索引擎实体推荐任务中所需处理的查询与实体的规模；(3) 传统推荐任务中对用户推荐的物品通常限定于同一个领域如商品或电影，而搜索引擎实体推荐则没有限定推荐实体的领域，它可能来自知识库中任何一个领域；(4) 在传统推荐任务中，用户对于物品的偏好信息通常能够显式地获取到，例如用户购买过某物品或观看过某电影的行为可以较为确定地表明用户对该物品或电影的喜爱，而在搜索引擎中用户对于实体的偏好信息则较难显式地获取到；(5) 在传统推荐任务中，由于被推荐的物品属于同一领域通常不需要给出具体的推荐理由，而搜索引擎中的实体推荐则需要给出具体的推荐理由以帮助用户更好地理解实体推荐结果.

面向搜索引擎的实体推荐存在的主要挑战及其对应研究任务如图 2 所示. 为了能更好地理解用户的搜索需求并准确地为用户推荐感兴趣的相关实体，一个完备的实体推荐系统需要包含三个子任务，分别为实体链接(entity linking)、实体推荐以及推荐理由生成. 其中实体链接旨在消除查询中实体指称的歧义并链接到知识库中无歧义的实体上，以获得与搜索查询对应的查询实体. 实体推荐旨在为查询实体寻找相关实体并排序生成推荐实体. 推荐理由生成则旨在为被推荐的实体集合以及单个实体生成推荐理由. 上述三个任务对应的技术挑战、研究现状以及解决方法将分别在第 2、3、4 节中进行详细介绍. 为更直观地进行说明，图 3 以搜索查询“美国总统奥巴马”为例，描述了实体推荐系统中不同模块的关系及工作流程.

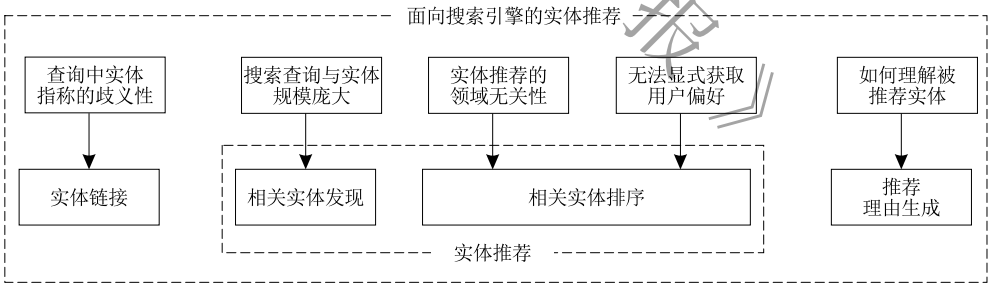

图 2　面向搜索引擎的实体推荐存在的主要挑战及其对应研究任务

实体链接的引入主要是针对以上所提的第一个挑战，其作用是将搜索查询中的实体指称消除歧义并链接到知识库中一个无歧义的实体上[11]. 实体链接不仅是实体推荐系统中必不可少的一部分[6,12]，也是知识库构建(knowledge base population)的重要环节[13]. 实体链接任务已经成为近年来的研究热点，国内外均有相关的实体链接评测如 TAC KBP[14]、ERD 2014[15]、NLPCC 2015[16]. 根据文本的形式，实体链接任务可以分为长文本实体链接与短文本(如 twitter[17-18]、搜索查询[12,19])实体链接. 从语言的角度，实体链接任务又可分为单语言实体链接与跨语言实体链接[20]. 实体链接任务通常包含三个子模块，分别是实体识别、候选实体获取与候选实体排序. 由于候选实体排序需要对实体指称及候选实体进行更深层的语义理解以计算它们的语义相似度，因而目前大部分对实体链接的研究都集中在候选实体排序阶段. 随着深度学习在自然语言处理领域不断取得进展[21]，实体链接的方法也不断发展，从依赖于人工构建特征[22]到利用神经网络从知识库和文本中自动学习特征[23-24]，实现对候选实体排序.

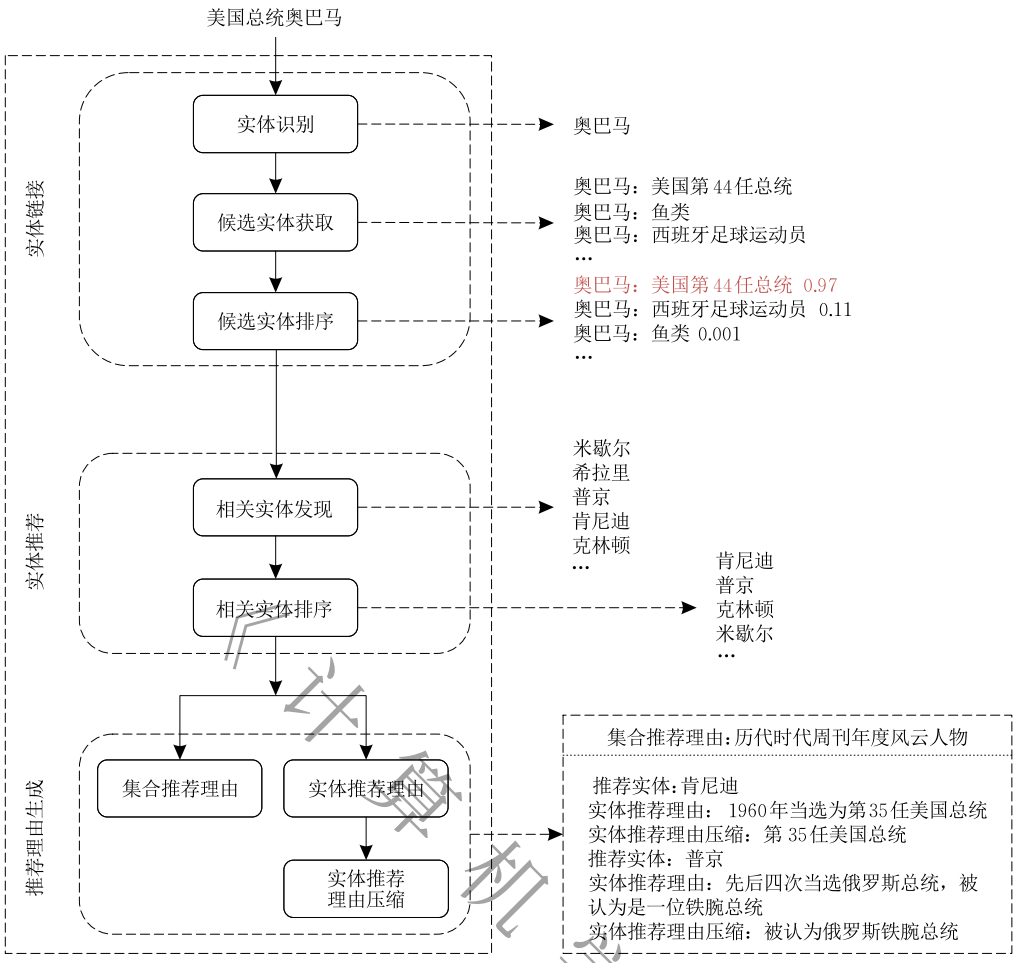

图 3 实体推荐系统示意图

实体推荐旨在为查询实体给出一系列推荐实体，它具有相关实体发现和相关实体排序两部分. 由于搜索查询与知识库的规模都很庞大，因此无法通过遍历的方式计算知识库中所有实体与查询实体的相关度来进行召回. 为了提高效率，相关实体发现模块为查询实体从知识库中召回一小部分最相关的候选实体. 由于实体推荐的领域无关性，在相关实体排序中需要尽可能引入更多领域无关的特征. 为了获取用户对实体的偏好信息，可以利用搜索日志中的用户点击信息. 按照是否利用当前查询的历史来搜索信息，目前的实体推荐方法可以被划分为上下文相关的方法[25-26]以及上下文无关[5-6,27-28]的方法. 其中上下文相关的方法由于考虑了用户的历史搜索，因而能够更好地对当前查询进行理解，使得给出的推荐结果与用户的信息需求更相关. 此外，按照是否考虑用户偏好信息，目前的实体推荐方法又可以被划分为个性化的方法[6,27-28]和非个性化的方法[5,25-26].

推荐理由生成旨在为被推荐实体集合以及单个被推荐实体生成推荐理由，分别为集合推荐理由和实体推荐理由. 集合推荐理由需要反映出被推荐实体集合与用户查询实体之间的关系，如图 1 中的集合推荐理由“相关人物”说明对应的被推荐实体均为与查询实体“奥巴马”相关的人物. 集合推荐理由的生成方法主要有基于标签的方法和基于模板的方法. 实体推荐理由通常可以分为两种：一种是关系型推荐理由，主要用来说明被推荐实体与查询实体的关系；另一种是亮点型推荐理由，旨在用简短的自然语言表达介绍被推荐实体的特点或独到之处. 例如在图 1 给出的推荐结果示例中，“第 42 任美国总统”主要介绍被推荐实体“威廉·杰斐逊·克林顿”，属于亮点型推荐理由. 而“92 年结婚并育有俩女儿”给出了被推荐实体“米歇尔·奥巴马”与查询实体之间的关系，则属于关系型推荐理由. 由于搜索引擎为了更好地展示实体推荐理由而对其字数进行了限制，因此实体推荐理由的生成需要分两步：(1) 为被推荐实体生成实体推荐理由，即一小段无字数限制的自然语言描述文本；(2) 对上述实体推荐理由进行压缩以使其符合搜索引擎要求的字数限制. 关系型实

体推荐理由和亮点型实体推荐理由的生成方法不同.目前关系型实体推荐理由的生成方法主要有基于模板的方法[29-30]和基于句子检索的方法[31-32],而亮点型实体推荐理由的生成方法主要是基于序列到序列学习的方法[33].对实体推荐理由进行压缩,既可以采用现有的句子压缩方法,也可以构建基于统计机器翻译的方法[34].

本文第 2 节介绍实体链接任务的定义、主要挑战、公开评测及数据集、任务划分以及各个子任务的研究现状;第 3 节首先介绍相关实体发现,然后介绍当前主流的实体推荐系统并分析其优点与不足;第 4 节介绍推荐理由生成任务的挑战及研究现状;第 5 节对实体推荐系统未来的研究方向进行展望;第 6 节为本文小结.

2　实体链接

查询中的实体指称通常具有歧义性,它可能指代知识库中的多个实体.例如实体指称"芝加哥"既可能指"芝加哥(城市)"也可能指"芝加哥(电影)".对于实体推荐任务而言,只有确定了查询中的实体指称在知识库中指代的实体才能够对其进行实体推荐.因此需要利用实体链接技术将查询中的实体指称消除歧义并链接到知识库中无歧义的实体上.

实体链接任务通常定义为给定一个知识库以及一段文本,识别出文本中的实体指称并将实体指称消除歧义链接到知识库中的对应实体上,如果该实体指称在知识库中没有对应的实体,则将其标记为 NIL[11].常用的外部知识库有 Wikipedia(维基百科)①[22,35]、DBpedia[10,36]、YAGO[37-38]、Freebase[9]等.实体链接任务的主要挑战在于处理名字的歧义性,歧义主要有以下两种[39]:(1)一个实体指称通常可以指代知识库中的多个实体,例如"苹果"既可以指"苹果(水果)"也可以指"苹果(公司)"、"苹果(电影)"等;(2)知识库中的实体通常具有多个名称如别名、简称等,例如美国歌手"泰勒·斯威夫特"常用的别名有"霉霉"、"TT"等.

根据文本类型可以将实体链接任务分为面向长文本(例如新闻、博客)的实体链接和面向短文本(微博、搜索查询)的实体链接.面向搜索查询的实体链接任务与面向长文本的实体链接任务相比更具挑战性,主要原因在于:(1)搜索查询通常较短,噪声大,且拼写错误和简写较多,缺乏充足的上下文;(2)面向搜索查询的实体链接要求更高的效率以及更低的空间占用.因此,直接将在长文本上表现较好的实体链接算法应用于搜索查询通常不能取得理想的效果.

面向长文本的实体链接评测有 ERD(Entity Recognition and Disambiguation Challenge)2014[15]和 TAC KBP(Text Analysis Conference Knowledge Base Population)2009－2018.ERD 2014 发布了长文本的实体链接任务,要求识别出网页中所有能够链接到知识库的实体指称.TAC KBP 2009 和 2010 的实体链接评测提供了一个由维基百科构建的知识库、待消歧的实体指称以及其所在的文档,若一个实体指称在知识库中存在对应实体则返回该实体,否则返回 NIL.TAC KBP 2011－2013 的实体链接评测包括了单语言实体链接和跨语言实体链接,另外要求将链接到 NIL 的实体指称进行聚类.TAC KBP 2014－2017 将实体链接任务定义为从文本中抽取实体指称并将实体指称链接到知识库,对链接到 NIL 的实体指称进行聚类.TAC KBP 2018 将实体链接任务的实体类型由 5 个扩充到 7309 个.Cucerzan[40]构建了一个用于实体链接的测试集合,其中包含 100 个不同主题的新闻故事,实体指称被链接到了维基百科中.Hoffart 等人[38]基于 CoNLL 2003 数据集人工构建了 AIDA② 数据集用于实体链接.AIDA 中包含 1393 篇新闻文章,文章中的每个实体指称都被人工标注出了在知识库 YAGO2 中对应的实体.其他常用的数据集还有 ACE[22]等.

面向短文本中搜索查询的实体链接评测主要有 ERD 2014[15]的短文本实体链接任务和 NLPCC(自然语言处理及中文计算会议)2015 发布的中文搜索查询中的实体识别和链接任务③,数据集有 YSQLE④(Yahoo Search Query Log To Entities).ERD 2014 的短文本任务中,给定一个搜索查询,要求利用所有可用的上下文给出所有合理的实体链接解释(entity linking interpretation).例如查询"total recall movie"存在两个合理的实体链接解释,"total recall"可能链接到知识库中的电影"total recall (2012)"或"total recall (1990)".NLPCC 2015 的实

① https://www.wikipedia.org/
② https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/
③ http://tcci.ccf.org.cn/conference/2015/pages/page05_evadata.html
④ https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

体识别和链接任务要求对于给定的中文搜索查询，识别出其中的实体并给出所有可能的实体链接解释. YSQLE 是专门针对搜索查询实体链接的数据集，其中包含了人工标注的搜索会话中对应到维基百科实体的链接，并提供了训练集合和测试集合. YSQLE 数据集中共包含了 2635 个查询，其中有 2583 个都标注了到维基百科实体的链接[19].

实体链接任务通常包含 3 个子任务[11]，分别是实体指称抽取(mention detection)、候选实体获取、候选实体排序. 由于知识库对实体的覆盖率有限且新实体不断出现，导致并非所有实体指称都能在知识库中找到对应的实体，因此还需要 NIL 判别，即判断一个实体指称是否在知识库中存在对应的实体.

实体指称抽取旨在从文本中抽取出所有可能被链接到知识库的实体指称. 实体指称抽取可以采用已有的命名实体识别的方法[7,41-42]，或者利用公开的命名实体识别工具如 Stanford NER Tagger[43]. 在抽取出实体指称之后，为提高效率需要从知识库中为每个实体指称生成可能的候选实体集合，之后可以通过计算集合中每个候选实体与实体指称的语义相似度等策略来确定实体链接的结果. 候选实体获取侧重于召回率，而候选实体排序更侧重于准确率. 相比于候选实体获取，候选实体排序更需要机器学习算法，因此当前大部分实体链接的研究工作侧重于候选实体排序的算法设计. 接下来本文将详细介绍候选实体获取、候选实体排序以及 NIL 判别. 由于实体指称抽取可以采用命名实体识别的方法，因此在本文中不再单独对其做详细介绍.

2.1 候选实体获取

知识库中通常包含千万级的实体，如果遍历整个知识库来计算每个实体与实体指称的语义相似度会非常耗时. 为了提高效率，需要为实体指称从知识库中找到最可能对应的实体集合.

有多种方式可以获取候选实体，常用的有以下三类：利用实体指称的上下文信息、利用维基百科构建词典以及利用搜索引擎.

(1) 利用实体指称的上下文信息. 实体指称的上下文包含了较丰富的信息，可以用来对实体指称进行扩展. 对缩写形式的实体指称进行扩展可以有效地降低实体指称的歧义性. Han 等人[44]通过人工制定启发式模板的方式从上下文中为缩写形式的实体指称找到候选实体，例如为实体指称 BBC 从上下文“The British Broadcasting Corporation (BBC) is a British public service broadcaster.”中抽取出候选实体“British Broadcasting Corporation”. 为了更好地从上下文中发现缩写形式实体指称的扩展形式，Zhang 等人[45]提出了一种基于有监督学习的方法. Gottipati 等人[39]利用命名实体识别工具从实体指称所在的上下文中识别出命名实体，制定规则并利用识别出的命名实体对实体指称进行扩展.

(2) 利用维基百科构建词典. 维基百科中的实体页面、重定向页面、消歧页面可以用来抽取实体名称与实体之间的映射关系[40].

实体页面是对一个实体进行描述的页面，通常包含结构化的信息盒子(infobox)以及描述文本. 从信息盒子中可以提取实体的别名，例如从图 4 所示的信息中可以直接提取出实体“凯蒂·佩里”的别名“水果姐”等. 在实体页面的首段描述文本，实体名称通常会以黑体表示，可以抽取出来作为该实体可能的实体名称. 实体页面中的超链接将一个实体指称链接到了一个实体上，因此可以用来提取实体名称与实体的对应关系. 通过对整个维基百科包含的实体页面中超链接的分析，还可以统计出实体名称链接到每一个目标实体的次数.

图 4 凯蒂·佩里的中文维基百科信息盒子

重定向页面通常只包含对一个实体页面的引用，别名可以被定向至实体页面，例如图 5 中“霉霉”被重定向到实体“泰勒·斯威夫特”.

图 5 维基百科的重定向页面示例

消歧页给出了某个实体名称可能对应的实体列表，例如图 6 给出了实体名称“苹果”对应的消歧页①，从中我们可以看到苹果可能链接到“苹果公司”等.

① https://zh.wikipedia.org/wiki/苹果_(消歧义)

图 6　维基百科消歧页示例

对于其他在线百科全书资源如百度百科①等也可以采用以上方法构建实体指称与实体的对应关系词典.

(3)利用搜索引擎. 利用现有的搜索引擎如百度、Google 等也可以获取实体指称的候选实体. Han 等人[44]将实体指称及其上下文送入 Google 搜索引擎,从搜索结果页中提取维基百科页面描述的实体作为候选实体.

召回率对于候选实体获取阶段很重要,因为正确的目标实体一旦没有被召回,则在后续的候选实体排序阶段也不能把实体指称链接到正确的实体上. 然而候选实体的数量也会影响最终实体链接的效果,过多的候选实体不仅会使候选实体排序阶段耗费更多时间,而且大量的可能完全不相关的候选实体也会给消歧带来挑战[46]. 提高消歧效率的一种方式是降低必须考虑的候选实体的数量[11]. 已有的大部分以召回率为驱动的候选实体获取策略都会增大候选实体的数量[47],因此如何在降低候选实体数量的同时保证较高的召回率也很值得研究.

Tan 等人[46]提出了一种候选实体获取方法,可以根据查询中的词过滤掉不相关的候选实体从而显著降低了候选实体的数量,其核心思想是从维基百科的文章中搜索与查询相似的句子,并且直接使用获取到的维基百科句子中人工标记的实体作为查询的候选实体. 与传统的候选实体获取方法相比,Tan 等人提出的基于句子搜索的候选实体获取方法产生的候选实体数量更少,且最终的消歧效果更好,这也说明了高质量的候选集合对于实体链接而言很重要.

2.2　候选实体排序

将实体指称链接到知识库中一个对应的实体通常可以视为对候选实体的排序问题. 已有的候选实体排序方法大致可以分为两种[20],分别是非联合(non-collective)的方法与联合(collective)的方法. 上下文中可能存在多个实体指称,非联合的方法每次只对一个实体指称进行消歧,而联合的方法还利用了上下文中实体指称相互之间的依赖关系对所有实体指称联合进行消歧.

非联合的候选实体排序方法主要考虑的信息有实体指称、实体指称的上下文、候选实体的名称、描述文档、热度等信息. 已有的非联合的候选实体排序方法设计了大量的丰富的特征集合,根据是否考虑实体指称所在的上下文,特征集合可以分为上下文无关的特征(如实体的热度以及实体名与实体指称的相似度等)与上下文相关的特征(如实体指称是否出现在候选实体的描述文档以及实体指称所在上下文与候选实体文档的语义相似度等). 例如 Dredze 等人[48]提出了五类特征,分别是名字变量有关的特征、维基百科特征、热度特征、文档特征以及特征组合. 其中名字变量有关的特征主要衡量的是实体指称与实体名的字面相似度、实体指称是否是实体的缩写或别名等,维基百科特征则是知识库属性方面的特征,热度特征为实体热度,文档特征主要是利用实体指称所在文档以及实体的描述文本抽取的特征. Zheng 等人[49]提出了三类特征,分别是表面(surface)特征、上下文(context)特征以及特殊特征. 其中表面特征衡量的是实体指称与实体名之间的字面相似度,上下文特征衡量的是实体指称与候选实体的上下文的相关性,特殊特征考虑了城市名以及实体指称和候选实体的类别.

同样的实体指称在不同的上下文下可能会链接到不同的实体,如何有效地利用实体指称的上下文信息对于实体链接任务非常关键[50]. 为了更好地学习到实体指称的上下文与候选实体之间的相关性,He 等人[23]提出了一种基于 DNN(Deep Neural Network)的方法利用 DA(Denoising Auto-Encoder)将实体指称与候选实体对应的文档映射到向量空间并计算它们的语义相似度,其网络结构如图 7 所示. Sun 等人[51]提出了一种用于实体链接的神经网络模型,并利用 CNN(Convolutional Neural Network)学习实体指称所在的上下文句子的语义向量表示. Fang 等人[52]提出了一种联合学习框架将知识库与文本联合映射到同一向量空间中以学习实体和词的

① https://baike.baidu.com/

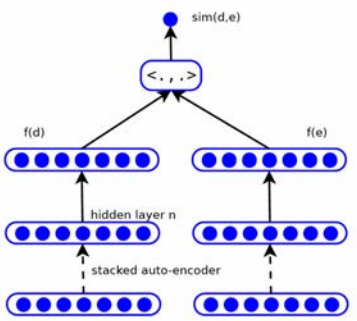

图 7 He 等人[23]提出的用于实体消歧的神经网络结构

低维连续的向量表示. Francis-Landau 等人[50]利用 CNN 学习实体指称的上下文与候选实体之间的语义对应关系(如图 8 所示). Yamada 等人[24]利用知识库的链接结构以及知识库的锚文本与上下文词对经典词向量学习模型 skip-gram[53-54]进行扩展,联合学习实体和词的向量表示. Gupta 等人[55]提出了一种实体链接的神经网络方法,利用实体的类型、无结构的描述文本与对应实体指称的上下文信息,通过组合的训练目标来学习实体的向量表示.

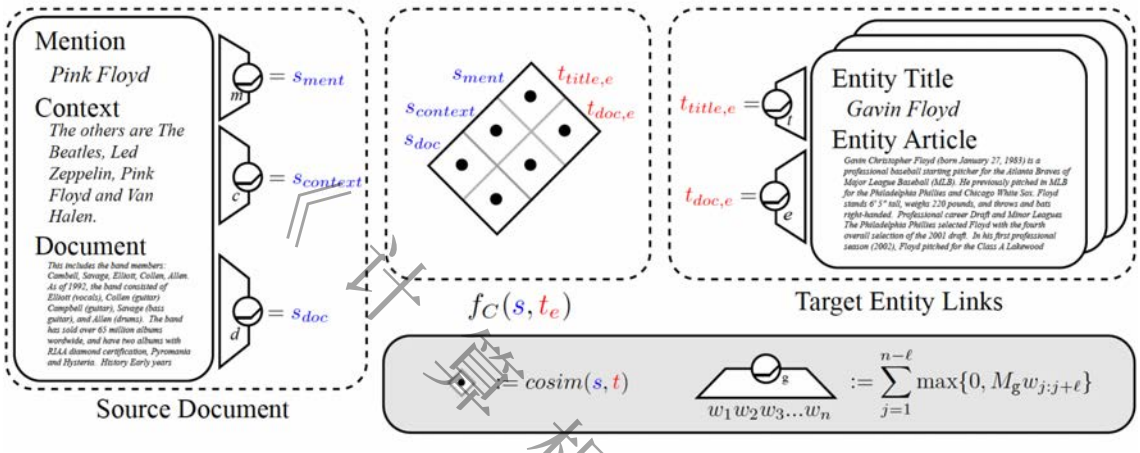

图 8 Francis-Landau 等人[50]提出的利用 CNN 构建特征

搜索引擎查询中的实体链接面临两个主要挑战:(1)搜索查询通常很短且包含噪声,没有足够的上下文来辅助实体指称的消歧;(2)搜索查询的实体链接通常需要在线处理,因此对速度的要求更高. 因此,在长文本中非常有效的实体链接方法不一定适用于搜索引擎的实体链接. Hasibi 等人[56]研究了查询中的实体链接,给定查询 q,要识别出一个实体链接解释集合 $I=\{E_1,\cdots,E_m\}$,每一个解释 E_i 是一个"实体指称-实体"的对应关系集合. 他们将该任务划分为候选实体排序和消歧两个子任务,候选实体排序的目的是从 q 中生成一系列"实体指称-实体"的排序列表,消歧的目的是将该排序列表作为输入并生成最终的实体链接解释集合 I. 通过在这两个子任务中分别应用有监督和无监督的方法,最终发现在候选实体排序阶段应用有监督学习方法,在消歧阶段应用无监督学习方法取得的效果最好. Blanco 等人[12]提出了一种非常快速并且空间效率高的概率模型来将查询链接到知识库的实体上去. 为了使算法快速并且空间效率高,该方法忽略了不同候选实体之间的依赖关系,并采用哈希和压缩的方法来减少内存占用. 此外,为了有效地利用查询中的上下文信息,该方法基于分布式语义表示计算查询和候选实体之间的相似度. 为了在候选实体发现阶段减少无关实体,Tan 等人[46]提出了一种非常简单且有效的搜索查询实体链接的方法,首先从维基百科中搜索与查询最相关的句子并直接将其中的标记实体作为该查询的候选实体,之后基于回归的框架采用丰富特征集合对候选实体进行排序.

在包含噪声较多的短文本中,依赖于 attention(注意力)的神经网络模型也不能总是找到正确的上下文线索,即便这些线索与目标实体的标题存在明显的字面重叠,而这种字面重叠的特征也难以用字符(character)级别的 CNN 学习到. 为了解决这一问题,Mueller 等人[57]构建了一个特征集合来表示实体指称的上下文与候选实体标题之间的字面重叠信息,并将这个特征集合融入到一个具有 attention 机制的神经网络实体链接模型中,模型的结构图如图 9 所示. 实验结果表明该特征的加入有效提升了神经网络模型的实体消歧效果.

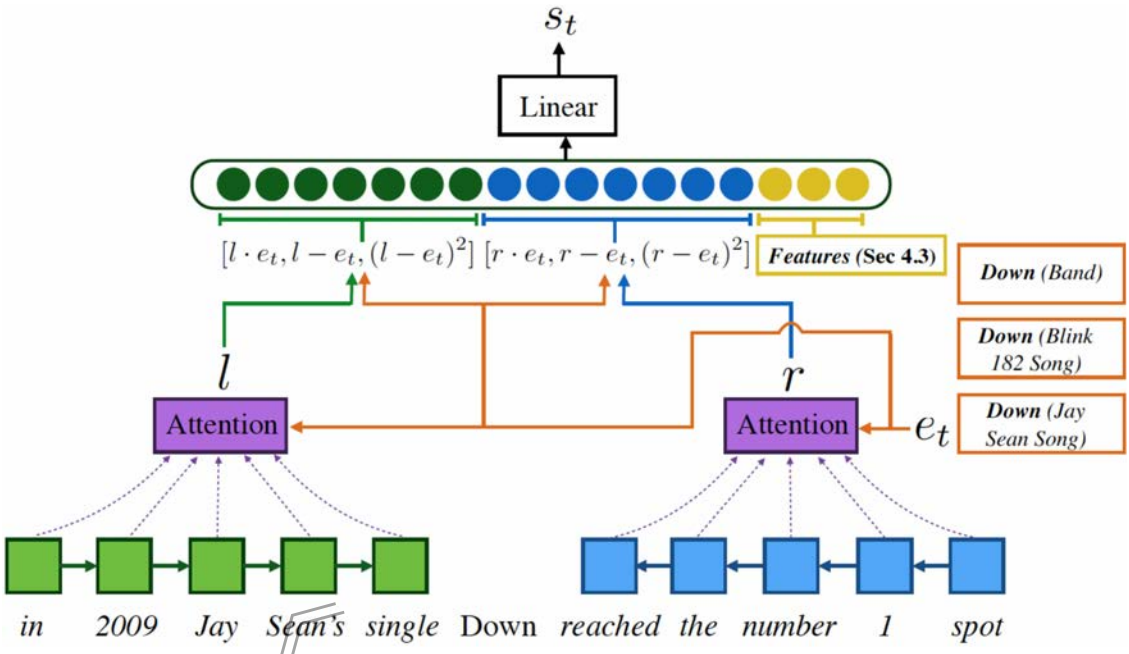

图 9　Mueller 等人[57]提出的实体消歧神经网络结构

神经网络模型既可以学习实体指称与候选实体的向量表示，也可以计算它们的语义相似度. 在训练阶段，神经网络模型的训练目标通常为使目标实体与实体指称的相似度得分高于错误的候选实体与实体指称的相似度得分[23,51]. 在预测阶段，给定实体指称以及一系列候选实体，可以根据模型计算出的相似度得分对候选实体进行排序.

与非联合的候选实体排序方法相比，联合的排序方法需要消歧上下文以及全局特征. 具体地，联合的候选实体排序方法不仅需要考虑单个实体指称与候选实体之间的语义相关性，还需要考虑出现在同一上下文中所有实体指称的预测实体之间的相关性. He 等人[58]提出了一种基于 stacking 的联合的候选实体排序方法，该方法由两层预测模型构成，底层是一个局部的预测模型 g_0，用于产生候选实体的初始排序结果，顶层是一个全局的预测模型 g_1，用于预测全局的排序结果. g_0 和 g_1 均为 LTR 模型，其中局部预测模型 g_0 的训练使用的是局部的特征，全局预测模型 g_1 的训练使用的是原始特征加上基于 g_0 的预测结果产生的全局特征. 消歧上下文为 g_0 对实体指称预测的 top k 的候选实体. 考虑到在长文档中不主要的实体可能只与文档中一小部分其他实体有关系，Globerson 等人[59]提出了一种基于 attention 的联合的实体链接方法.

2.3　NIL 判别

知识库通常无法覆盖出现在互联网文本中的所有实体尤其是新出现的实体，因此并非所有实体指称都可以被链接到知识库的实体上. 不能被链接到知识库的实体指称被返回一个特殊实体 NIL[11]. 目前对实体指称是否链接到 NIL 的判别方法可以分为以下几种：

(1) 最简单的方式是设置阈值判断候选实体排序阶段排序第一的候选实体是否为最终目标实体. Gottipati 等人[39]利用 NER 工具识别出实体指称的实体类别，然后从候选实体集合中找出经过候选实体排序阶段排序最靠前且与实体指称属于同一类别的候选实体，若该候选实体在候选实体排序阶段的得分高于某个预设的阈值则将其作为实体链接结果，否则将实体指称链接到 NIL. Han 等人[44]设置了一个阈值来判断实体指称是否需要链接到 NIL，若所有候选实体与实体指称计算出的相似度均小于该阈值，则将实体指称链接到 NIL. 相似地，Nie 等人[60]根据开发集合设置了一个 NIL 阈值. 虽然通过设定阈值的方式来判断 NIL 很简单也不需要引入模型和特征，然而人工设定阈值比较困难，并且为所有样本应用同一个阈值也并不是特别合理.

(2) 把 NIL 作为一个特殊实体加入候选实体集合，若在候选实体排序阶段 NIL 排序最靠前则判定该实体指称应当被链接到 NIL. Han 等人[61]将 NIL 看作一个虚拟实体加入到知识库中并将 NIL 实体与其余实体一同对待，若计算得出的由 NIL 实体生成实体指称的概率比知识库中其他实体生成实体指

称的概率更高，则将实体指称链接到 NIL. Dredze 等人[48]将 NIL 作为一个特殊的候选实体和其他候选实体一起进行排序，并且为排序模型专门设计了一些与 NIL 相关的特征，例如是否存在候选集合中的一个实体其名称与实体指称相匹配. 将 NIL 看作一个特殊候选实体的优点是既避免了人工设置阈值又可以将与 NIL 判别相关的信息作为特征引入到排序模型中.

(3)将 NIL 判别看作二分类任务，利用分类器判断候选实体排序阶段给出的排序第一的候选实体是否合理，若合理则将该候选实体作为最终实体链接的结果，否则该实体指称将被链接至 NIL. Zheng 等人[49]和 Zhang 等人[45]用 SVM 分类器判断排序第一的候选实体是否为真正的目标实体，在分类器中采用的特征大部分与候选实体排序阶段所用的特征相同.

除了简单地判断实体指称是否指向 NIL，TAC KBP 的实体链接评测任务自 2011 年起要求将所有指向 NIL 的查询(实体指称及其上下文)进行聚类. 将指向 NIL 的查询聚类对知识库的扩充非常有帮助，被聚到同一类别的实体指称及其上下文代表了一个新实体相关的信息. 对 NIL 型查询进行聚类可以借助已有的聚类算法. Taylor 等人[62]基于简单的子串匹配的方法进行 NIL 聚类. Graus 等人[63]将查询对应的源文档表示为 TF×IDF 向量，然后采用层次聚类的方法基于这些向量表示进行聚类. 由于对 NIL 型查询的聚类与实体推荐任务的相关度并不是很高，因此在本文中不做更详细的介绍.

2.4 面向实体推荐任务的实体链接

对于当前大部分实体推荐系统而言，其输入为一个仅包含实体指称的搜索查询 q. 仅通过 q 本身，无法获取任何对 q 的消歧有帮助的上下文信息，因此目前大部分实体推荐系统[5,27]基于热度 $p(e|q)$ 对 q 进行实体链接，这样的结果导致 q 只能够被链接到最热的实体上.

实际上，q 所在的搜索会话中的历史查询及其点击信息可以看作 q 的上下文信息[6,26]. 例如，一个用户在搜索了“香蕉牛奶”后再搜索“苹果”，那么当前的查询“苹果”很大概率应该被链接到知识库的实体“苹果(水果)”上. 若用户在搜索了“华为 P20”之后再搜索“苹果”，那么当前查询“苹果”更大概率应该被链接到“苹果(公司)”.

将 q 所在的搜索会话中的历史查询及其点击信息看作上下文，可以利用已有的实体链接方法[12]将 q 链接到实体. 在对 q 的候选实体排序时，既可以采用非联合的候选实体排序方法[49]只对 q 进行消歧，也可以利用联合的候选实体排序方法[52]对 q 所在搜索会话中所有的实体指称进行消歧.

q 的上下文不同于传统实体链接任务中实体指称的上下文. 传统实体链接任务中实体指称的上下文通常为无结构的短文本或长文本，而 q 的上下文是有层次结构的，由历史查询及其点击信息构成. 历史查询的点击信息对查询中的实体指称消歧具有非常重要的作用[26]. 例如，如果用户搜索了“苹果”并点击了标题为“Apple(中国)-官方网站”的网页，则该“苹果”很大概率指的是知识库中的“苹果(公司)”. 如果用户搜索“苹果”并点击了标题为“苹果的家常做法”的网页，则该“苹果”很大概率指的是知识库中的“苹果(水果)”. 在人工设计特征或者利用神经网络自动学习特征时应当考虑到 q 的上下文的这一特点.

2.5 实体链接方法总结

综上，我们认为目前的实体链接算法具有很强的表示能力，尤其在深度神经网络的框架基础上，配以充足有指导信息便可以训练获得较高性能的实体链接系统. 然而，在很多新的应用场景中没有足够的有指导数据供模型训练，如何在这种资源匮乏的情景下训练模型是一个在实用中遇到的挑战. 此外，当前的实体链接算法大多假设训练和测试数据的分布相同或相似，这样的系统很难应付对抗样本，即数据分布不同于训练数据的样例，如何增强系统的鲁棒性也是一个非常值得研究的方向. 再次，当前实体推荐系统所处理的搜索查询只包含一个实体指称而没有其他任何上下文信息，单纯从这个实体指称本身无法推断用户究竟想搜索的是哪一个实体. 而用户在同一个搜索会话中的搜索查询之间具有一定的任务相关性，因此可以借助于本次搜索会话中的历史查询及其点击信息对当前查询进行实体链接[26]. 历史查询中的实体指称本身也是有歧义的，因此可以采用联合的方法同时对历史查询中的实体指称和当前查询的实体指称进行实体链接.

3 实体推荐

在搜索引擎中，实体推荐系统的目标是为给定用户 u 推荐与其输入的查询 q 相关的一系列实体 E_{uq}，即 $R(u,q)\mapsto E_{uq}$. 实体推荐主要由相关实体发现与相关实体排序两部分构成. 具体地，给定一个用

户 u 输入的查询 q 所对应的查询实体 e_q，实体推荐任务的目标是首先获取一系列与 e_q 相关的实体 $\mathcal{R}(e_q)=\{e_1,e_2,\cdots,e_n\}$，然后学习出一个对候选实体与用户信息需求之间的匹配度进行打分的函数，根据 $\mathcal{R}(e_q)$ 中各个候选实体的得分对其进行排序后，再基于排序结果获得实体推荐结果 E_{uq}. 虽然也存在其他实体推荐任务，例如面向用户的个性化实体推荐[64-65]，但该任务要解决的问题是为给定用户 u 推荐其感兴趣的实体集合 E_u，即 $R(u)\mapsto E_u$. 虽然任务名称相同，但从定义中可以看出，面向搜索引擎的实体推荐与面向用户的个性化实体推荐之间存在显著差异. 此外，在搜索引擎中，如果只考虑用户兴趣，而不考虑用户输入的查询，则会导致推荐出的实体虽与用户相关，但与用户的信息需求却毫不相关，从而不可避免地会损害用户的搜索体验. 本文重点介绍面向搜索引擎的实体推荐任务.

3.1　相关实体发现

在实体推荐系统中，相关实体发现任务的目标是获取与查询实体 e_q 相关的候选实体并按照与 e_q 的相关度进行排序，得到 e_q 的相关实体集合 $\mathcal{R}(e_q)=\{e_1,e_2,\cdots,e_n\}$. TREC 2009 也存在一个相关实体发现任务[66-67]，其定义为：给定一个实体 e_q（包含实体名称与实体主页）、目标实体类别 T 以及含有实体关系描述的自由文本 n_{ar}，找出与 e_q 具有 n_{ar} 中所述关系且类别为 T 的目标实体. 虽然这两个任务的名称相同，但从定义中可以看出，二者间存在显著差异. 本文重点介绍实体推荐系统中相关实体发现任务.

实体推荐系统中的相关实体发现，按照数据源与实体间相关度计算方法的差异，目前的方法大致可以分为以下 3 类：

(1) 基于知识图谱的方法. 知识图谱是一种存储实体及其关系的集中式知识库. 知识图谱中构建的实体关系都是已知的事实. 因此，在知识图谱中与 e_q 间存在关系的其他实体，都可以被直接抽取出来作为 e_q 的相关实体集合（记为 $\mathcal{K}(e_q)$）. 该思路被广泛应用于目前的各种实体推荐方法中[5-6,27-28]. 通过这种方法获取的相关实体，可以基于以下方法计算 e_q 与候选实体 $e_c\in\mathcal{K}(e_q)$ 间的相关度 $rel(e_q,e_c)$. 首先，如果 e_q 与 e_c 在知识图谱中存在直接关系，则 $rel(e_q,e_c)$ 可按下式进行计算：

$$rel(e_q,e_c)=\begin{cases}1, & e_c\text{与}e_q\text{在知识图谱中有直接关系}\\ 0, & \text{除上述情况以外}\end{cases} \tag{1}$$

其次，对于在知识图谱中无直接关系的实体，可以基于图路径计算两个实体间的相似度，作为二者相关度得分[27,68]. 例如，如果两个实体间由给定路径类型 $\mathcal{P}$（例如“电影-演员-电影”）的路径所链接，则 $rel(e_q,e_c)$ 可按下式进行计算：

$$rel(e_q,e_c)\approx sim_{\mathcal{P}}(e_q,e_c)=\frac{2\times|\{p_{e_q\to e_c}:p_{e_q\to e_c}\in\mathcal{P}\}|}{|\{p_{e_q\to e_q}:p_{e_q\to e_q}\in\mathcal{P}\}|+|\{p_{e_c\to e_c}:p_{e_c\to e_c}\in\mathcal{P}\}|} \tag{2}$$

上述公式中的 $p_{e_q\to e_c}$ 表示两个实体间的一个路径. 此外，还可以根据知识图谱中两个实体的属性或描述文本间的相似度[6,27]来计算两个实体间的相关度得分. 例如，可以计算出两个实体对应的百科文章间的内容相似度[6]，将其用于衡量二者之间的相关度. 该方法使用潜在狄利克雷分布（Latent Dirichlet Allocation，简称 LDA）[69]对每篇百科文章进行建模，再用训练得到的 LDA 模型获得文档 d 所对应的主题向量表示 $\boldsymbol{v}_d$. 然后，按照如下公式计算出实体 e_q 与 e_c 所分别对应的百科文章 d_q 与 d_c 的主题向量 $\boldsymbol{v}_{d_q}$ 与 $\boldsymbol{v}_{d_c}$ 之间的相似度，即可获得这两个实体间的相关度得分：

$$rel(e_q,e_c)\approx sim(e_q,e_c)=\cos(\boldsymbol{v}_{d_q},\boldsymbol{v}_{d_c})=\frac{(\boldsymbol{v}_{d_q})^{\mathrm{T}}\boldsymbol{v}_{d_c}}{\|\boldsymbol{v}_{d_q}\|\|\boldsymbol{v}_{d_c}\|} \tag{3}$$

(2) 基于搜索日志的方法. 由于知识图谱中的实体及其关系信息往往不太完备，基于知识图谱获得的相关实体集合往往覆盖率有限. 为缓解这一问题，可以基于搜索会话共现来补充完善 e_q 的相关实体集合. 使用搜索引擎的一部分用户会在同一个搜索会话中多次搜索不同的查询[70]，用户的这些搜索行为所积累起来的搜索日志是一种有效的挖掘实体及其关系的数据来源，因为这些信息能够有助于发现一些实体间存在的潜在关联关系. 具体地，可以将与 e_q 在同一搜索会话中多次共现，且共现次数高于某个阈值的实体抽取出来，作为 e_q 的候选相关实体集合（记为 $\mathcal{S}(e_q)$）. 该思路在目前的各种实体推荐方法中也被广泛采用[5-6,27-28]. 给定查询 e_q 与候选实体 $e_c\in\mathcal{S}(e_q)$，可以基于互信息来估算二者之间的相关度得分：

$$rel(e_q,e_c)=\frac{PMI(e_c,e_q)}{\sum\limits_{e_c'\in S(e_q)}PMI(e_c',e_q)} \tag{4}$$

上述公式中的 $PMI(e_c,e_q)$ 定义如下：

$$PMI(e_c,e_q)=\log\frac{cnt(e_c,e_q)}{cnt(e_c)\cdot cnt(e_q)} \tag{5}$$

在上述公式中，$cnt(e_c,e_q)$为e_c与e_q共同出现过的总搜索会话个数，而$cnt(e_q)$为含有e_q的总搜索会话个数.

(3)基于网页文档的方法.基于搜索会话共现信息补充e_q的相关实体，这种方法虽然有效但却严重依赖于搜索日志数据对实体的覆盖率.也就是说，该方法只能局限于使用现有搜索会话日志中出现过的实体共现信息，帮助发现与e_q相关的其他实体.由于新出现的实体往往缺乏共现信息，因此很难通过基于搜索日志的方法为新实体找到相关实体.为缓解这一问题，可以基于给定的网页文档，首先从中抽取出与e_q共现过的所有实体(记为$\mathcal{D}_r(e_q)$)，然后基于$\mathcal{D}_r(e_q)$中各候选实体与e_q间的相关度对其进行排序，最后将相关度得分高于某个阈值的实体抽取出来，作为e_q的候选相关实体集合(记为$\mathcal{D}(e_q)$).基于网页文档的方法主要可以分为三类：基于分类体系、基于链接结构以及基于文本内容.

首先，可以基于两个实体在给定实体分类体系(例如维基百科中的分类体系)中的相对位置(包括分类层次深度、路径长度等)计算二者之间的相关度[71-73].由于实体分类体系往往很难覆盖所有领域，而构建特定领域的分类体系又是一件复杂与耗时的工作.因此，该算法的通用性较大程度受制于分类体系的覆盖率与完备度.

第二类方法主要基于实体之间的链接结构计算相关度[74-77].这类方法主要基于随机游走算法挖掘实体之间的引用结构，从而能够发现一些实体间潜在的深层次相关关系.这类方法主要采用链接分析算法对维基百科等文档资源中的各个实体间存在的链接进行分析后挖掘出实体间的关系.不足之处在于这类算法迭代更新的代价较大，而在搜索情景下实体变化较快，导致其可用性有限.

第三类方法主要基于文本内容计算实体相关度.这类方法主要基于实体间共享的某些文本信息计算实体之间的相关度.计算时主要采用的信息包括百科文章中重叠的词与短语[78]、重叠的超链接[79]、事先从文章中抽取出的关键短语[80]或者从百科文章中学习得到的实体概念向量[81-82]与分布式表示[83-84].与分类体系相比，网页文档的覆盖率与完备度更高，且获取起来更容易.而与基于链接结构的方法相比，这类方法可以借助的信息更多、灵活性更高且无需全局迭代.

由于采用单个方法很难获得理想的结果，因此往往通过对上述三种方法得到的相关实体集合进行合并，获得与e_q相关的实体集合$\mathcal{R}(e_q)$：

$$\mathcal{R}(e_q)=\mathcal{K}(e_q)\bigcup\mathcal{S}(e_q)\bigcup\mathcal{D}(e_q) \tag{6}$$

3.2 相关实体排序与推荐

相关实体排序是实体推荐任务的核心部分.具体地，给定一个用户u所输入的查询实体e_q及其相关实体集合$\mathcal{R}(e_q)=\{e_1,e_2,\cdots,e_n\}$，相关实体排序任务的目标是学习出一个对实体与用户信息需求之间的匹配度进行打分的函数，然后按照匹配度得分对$\mathcal{R}(e_q)$中的各个实体进行排序.

首先，在排序目标上，可以采用相关度或兴趣度作为匹配度的优化目标.相关度主要基于人工标注的候选实体相关性等级进行建模，而兴趣度则基于真实的用户实体点击数据进行建模.因此，按照所采用的排序目标，目前的实体推荐方法大致可以被划分为两类：基于相关度的实体推荐方法[5]与基于兴趣度的实体推荐方法[6,25-28].

其次，在对相关实体进行排序时，大部分实体推荐方法都只考虑了候选实体与用户u所输入的查询实体e_q之间的关系，而未考虑当前搜索会话中的上下文信息(即用户在同一搜索会话中输入的历史查询及其对应的点击信息)与用户偏好信息.上下文信息与用户偏好信息对于更准确地理解用户信息需求具有重要作用.具体地，上下文信息有助于更好地理解用户查询背后的搜索意图，对于提升实体推荐结果的相关性具有重要作用.而用户偏好信息有助于更好地理解用户u对不同实体的偏好程度，对于提升实体推荐结果的个性化程度具有重要作用.因此，可以按照是否考虑上述两种信息对目前的实体推荐方法进行分类.按照是否考虑上下文信息，可分为上下文无关(context-insensitive)的实体推荐方法[5-6,27-28]与上下文相关(context-aware)的实体推荐方法[25-26].而按照是否考虑用户偏好信息，可分为个性化实体推荐方法[6,27-28]与非个性化实体推荐方法[5,25-26].

此外，一些实体推荐方法受限于领域知识与特定领域的实体属性及实体关系.因此，这些方法[27-28]需要先明确领域，才能为给定领域与类别下的查询进行实体推荐.

表1对目前的实体推荐方法在上述因素上的差异进行了汇总.下面分别从排序目标、排序方法以及实体推荐模型等方面对各个实体推荐方法进行介绍与分析.

表 1　不同实体推荐方法对比

	查询	搜索会话	用户偏好	领域无关	排序目标
Spark[5]	√	×	×	√	相关度
PRM-KNN[27]	√	×	√	×	兴趣度
TEM[28]	√	×	√	×	兴趣度
LTRC[6]	√	×	√	√	兴趣度
CF[25]	√	√	×	√	兴趣度
ER-C-MT[26]	√	√	×	√	兴趣度

Blanco 等人[5]提出了一种基于排序学习的实体推荐方法，根据实体间的相关度排序结果进行推荐. 该方法解决的任务是估算给定查询实体 e_q 与候选相关实体 e 间的相关度，即估算概率 $P(e|e_q)$. 该方法基于知识图谱、搜索日志、社交网站等抽取特征，并基于 5 个相关性等级对 e_q 与 e 间的相关度进行人工评分后，基于人工标注的相关度训练排序学习模型. 该方法的主要不足在于只考虑实体间的相关度，而未考虑用户对候选实体的兴趣度，因此可能无法有效地满足用户的实际信息需求.

Bi 等人[28]与 Yu 等人[27]提出了基于搜索日志与知识图谱的实体推荐方法，为特定领域（例如电影、人物等）的查询进行个性化实体推荐. 这些方法依赖于众多有关领域的特征（例如电影类型、用户查看过的电影导演等）. 因此，这些方法需要先明确领域，才能为给定领域与类别下的查询进行相关实体推荐. 具体地，给定一个由用户 u 所输入的类别为 T 的查询实体 e_q^T，这些方法的目标是从搜索日志与知识图谱中抽取出一系列与 u，e_q^T 以及候选实体 e 相关的特征，并基于众多特征学习出一个能够估算用户 u 在输入 e_q^T 时对 e 的兴趣度的打分函数 $f(u,e_q^T,e)$，从而根据兴趣度对候选相关实体集合进行排序. 实验结果表明，实体点击率是所有特征中最有效的特征，对于提升实体推荐系统的效果具有至关重要的作用. 这一结论也说明，在构建面向搜索引擎的实体推荐系统时，如果只考虑实体间的相关度，而不考虑用户对实体的兴趣度，很难有效满足用户的实际信息需求. 这些方法的主要不足是依赖于知识图谱中的实体属性与实体关系，而知识图谱中的实体信息往往很难完备并保持及时更新，从而不可避免地会影响这些方法的实际应用效果. 此外，由于这些方法依赖于领域相关的特征，因此只能为明确类别的查询进行实体推荐，而无法通过一个模型为所有类别的查询进行实体推荐. 由于用户在搜索引擎中输入的查询的类别是开放领域的，如果要为所有查询进行实体推荐，则需要按照类别逐一构建不同的实体推荐模型. 上述限制降低了这些推荐方法的通用性，也增加了这些方法在搜索引擎中实际落地应用的难度.

Huang 等人[6]提出了一种由相关实体发现与相关实体排序两部分所构成的实体推荐框架，根据查询实体与用户偏好为用户推荐相关且带有惊喜度（serendipity）的实体. 具体地，给定一个由用户 u 所输入的查询实体 e_q，该方法的目标是抽取出一系列与 e_q 相关的实体 $\mathcal{R}(e_q)=\{e_1,e_2,\cdots,e_n\}$，并学习出一个能够估算用户 u 对候选相关实体 e 的兴趣度的打分函数 $f(u,e_q,e)$，从而根据兴趣度得分对 $\mathcal{R}(e_q)$ 中的实体进行排序. 为了更好地提升实体推荐的惊喜度，该框架针对惊喜度三要素（相关度、意外度以及兴趣度）设计了三组特征. 实验结果表明，该方法在实体推荐效果上显著优于多个稳健的基线方法. 消融实验（ablation study）的结果表明兴趣度特征是其中最有效的特征，且意外度特征能够显著帮助提升实体推荐的效果. 在百度搜索引擎上进行的在线对照实验的结果表明，该方法产出的实体推荐结果还能显著提升用户参与度. 但不足之处在于该方法只考虑了用户与查询实体，而未考虑同一搜索会话中的上下文信息.

Fernandez-Tobias 等人[25]提出了一种基于记忆（memory-based）的实体推荐方法，根据搜索会话为用户推荐相关实体. 具体地，给定一个搜索会话 s 与候选相关实体 e，该方法的目标是估算二者相关的概率 $P(e|s)=\sum_{\bar{e}\in E(s)}P(e|\bar{e})P(\bar{e}|s)$. 该公式中的 $E(s)$ 为搜索会话 s 中用户点击的实体集合，$P(e|\bar{e})$ 为实体 e 与 $\bar{e}$ 间的相似度，而 $P(\bar{e}|s)$ 为 $\bar{e}$ 与 s 的相关度. 该方法基于最近邻协同过滤推荐算法[85-86]，只依赖于搜索日志中用户的过往搜索行为，因此不受限于特定领域. 但不足之处在于该方法完全依赖于用户行为数据，因此不可避免地会受制于数据稀疏与冷启动问题，尤其是缺乏用户行为数据的长尾、冷门查询以及新实体. 此外，当候选实体集合中出现新实体时，必须重新计算实体相似度并重训模型. 由于现实场景中新实体会持续不断出现，从而制约了该方法在搜索引擎中的实际落地应用.

为了提供与用户信息需求更相关的实体推荐结果，需要对隐含在用户查询背后的搜索意图进行更好的理解. 为此，Huang 等人[26]提出了一种基于深度多任务学习的上下文相关的实体推荐模型，以上下文相关的文档排序作为辅助任务，并借助于搜索会话中的历史信息（如前序查询序列）来更好地理解

当前查询的搜索意图，从而为用户推荐与其信息需求更相关的实体. 具体地，给定一个查询 q_t（除当前搜索会话中的首个查询外，即 $t\neq 0$）与当前搜索会话中的前序查询序列构成的上下文 $c=q_1,q_2,\cdots,q_{t-1}$，该方法的目标是通过神经网络学习出一个表示函数，用于将 q_t 与 c 转化为向量表示 $\boldsymbol{v}_m$，并将候选相关实体 e 转化为向量表示 $\boldsymbol{v}_e$，然后按如下公式计算出 $\boldsymbol{v}_m$ 与 $\boldsymbol{v}_e$ 之间的相似度：

$$P(e|c,q_t)=\cos(\boldsymbol{v}_e,\boldsymbol{v}_m)=\frac{\boldsymbol{v}_e^{\mathrm{T}}\boldsymbol{v}_m}{\|\boldsymbol{v}_e\|\,\|\boldsymbol{v}_m\|} \tag{7}$$

最后再按照相似度得分对所有候选实体进行排序. 图 10 显示了基于深度多任务学习的上下文相关的实体推荐模型的网络结构. 从中可以看出，用于学习查询与上下文表示的网络层（图中下方部分）在两个任务间是共享的，而其他网络层则是任务相关的（图中上方部分）. 共享表示通过优化多任务学习目标进行学习，学习得到的共享表示 v_s 概括了当前查询与所有前序查询的信息. 而实体与文档的表示则分别通过优化任务相关的目标进行学习. 实验结果表明，引入上下文信息能够显著提升实体推荐的效果，而且基于多任务学习的方式训练模型能够带来进一步的效果提升. 但不足之处在于该方法只考虑了查询实体与同一搜索会话中的前序查询，而未考虑搜索会话中的其他上下文信息与用户偏好信息. 例如，用户在当前搜索会话与历史搜索中的点击信息. 此外，该方法在对实体进行表示学习时采用的是实体编号，因此在将模型应用到搜索引擎时，不可避免地会面临新实体的冷启动问题.

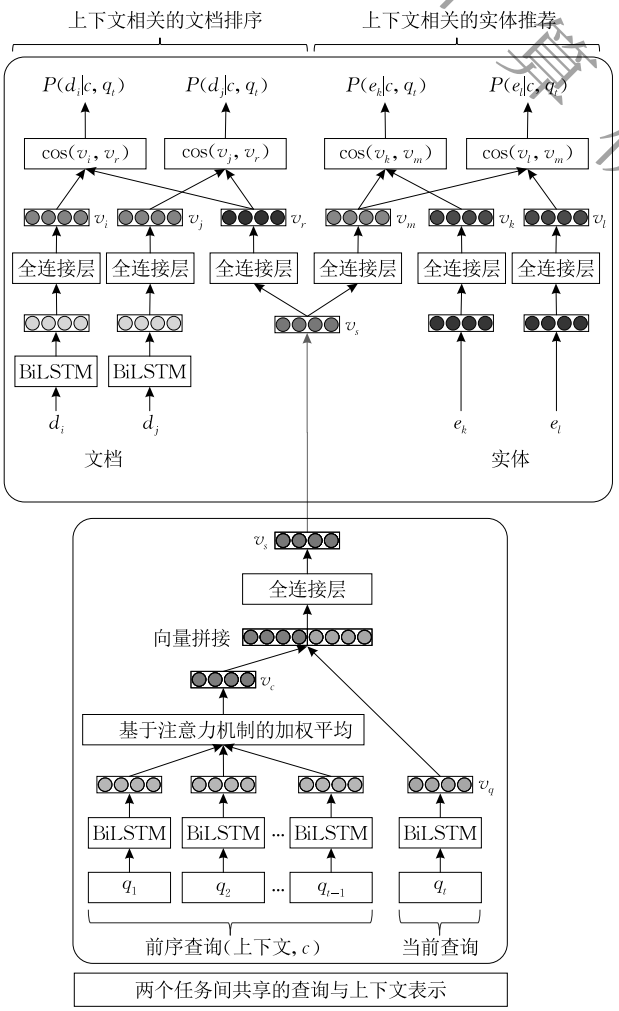

图 10 Huang 等人[26]提出的基于深度多任务学习的上下文相关的实体推荐模型的网络结构

3.3 实体推荐评价指标

在对实体推荐方法的效果进行评价时，目前的方法主要可分为以下两类：离线评价与在线评价. 离线评价旨在对不同实体推荐方法给出的实体推荐结果的质量进行评价. 而在线评价则旨在通过真实的大规模用户行为数据对不同实体推荐结果的用户参与度进行评价.

在实体推荐任务中广泛采用的离线评价指标主要包括以下两种：折扣累积增益 *DCG*（Discounted Cumulative Gain）[87] 与平均排序倒数 *MRR*（Mean Reciprocal Rank）. *DCG* 与 *MRR* 是信息检索中用于衡量排序质量的评价指标，被广泛应用于评价搜索结果的相关性. 为了对比各排序模型在不同排序位置上的效果，通常会基于不同检查点上的 *DCG* 值与 *MRR* 值分别对各排序结果的质量进行比较. 累计在给定排序位置 p 处的 *DCG* 的计算方法如下：

$$DCG_p=\sum_{i=1}^{p}\frac{2^{rel_i^{uq}}-1}{\log_2(i+1)} \tag{8}$$

在上述公式中，$rel_i^{uq}\in\{0,1\}$ 是用户 u 输入的查询实体 e_q 的相关实体排序结果中位于排序位置 i 的实体的二元相关性（1 为相关，0 为不相关）. 整体 *DCG* 值由给定测试集中的所有测试样本的 *DCG* 经过求平均值后计算得出. 而前 p 个排序结果的平均排序倒数的计算方法如下：

$$MRR_p=\frac{1}{|\mathcal{T}|}\sum_{j=1}^{|\mathcal{T}|}\frac{1}{rank(e_j^{uq})} \tag{9}$$

在上述公式中，e_j^{uq} 为第 j 次测试时，与用户 u 所输入的查询 e_q 所对应的标准实体（ground truth entity），$rank(e_j^{uq})$ 则表示由给定排序方法所确定的 e_j^{uq} 的排序位置. 如果标准实体 e_j^{uq} 不在给定的前 p 个排序结果中，则 $rank(e_j^{uq})$ 取值为 ∞. $\mathcal{T}$ 为测试集合.

在线评价旨在评价用户对推荐结果的参与度. 为了量化并比较不同实体推荐方法的用户参与度，

常用的方法是比较实体推荐结果的点击率 *CTR* (Clickthrough Rate). *CTR* 是一种用于评价在线服务效果的有效评价指标，例如可用于评价推荐系统效果[88]、评价搜索广告效果[89]以及评价网页搜索结果页面功能[90]. 在搜索引擎公司中，基于大量在线对照实验后获得的真实用户数据进行效果评价的方式已被广泛采用[91]. 而在线对照实验通常以 *A*/*B* 测试①的方式进行. 通过比较在线对照实验中各个实体推荐方法对应的 *CTR* 值，即可对不同实体推荐方法的用户参与度进行对比评价. *CTR* 值越大，则表明用户参与度越高. *CTR* 值的计算方法如下：

$$CTR=\frac{\sum_{e_q\in\mathcal{T}}\sum_{e_c\in\mathcal{R}(e_q)} click(e_q,e_c)}{\sum_{e_q\in\mathcal{T}}\sum_{e_c\in\mathcal{R}(e_q)} impression(e_q,e_c)} \tag{10}$$

在上述公式中，$\mathcal{R}(e_q)$ 为 e_q 的推荐实体集合，$click(e_q,e_c)$ 为用户搜索 e_q 时对实体 e_c 进行点击的总次数，$impression(e_q,e_c)$ 则为用户搜索 e_q 时，展现过实体 e_c 的页面的总数.

采用离线评价的方式，只能评价推荐结果的质量，无法评价真实用户对一个实体推荐方法所给出的推荐结果的参与度. 与离线评价的方式相比，由于在线评价基于真实用户的大规模行为数据，因此在线评价相较于离线评价其结果往往与模型上线后的实际应用效果更吻合. 在目前的实体推荐方法中，大部分工作[25-28]都只采用了离线评价的方式，只有少部分工作[5-6]既采用了离线评价的方式，又采用了在线评价的方式.

3.4　实体推荐方法总结

面向搜索引擎的实体推荐任务主要面临以下三个挑战：搜索查询与实体规模庞大、领域无关、无法显式获取用户偏好. 首先，搜索查询与实体规模庞大主要会导致计算量与冷启动两方面的问题. 为解决该挑战，Huang 等人[6]提出了一种由相关实体发现与相关实体排序两部分构成的实体推荐框架，能够有效解决上述问题. 一方面，在相关实体发现阶段，通过对基于知识图谱、搜索日志与网页文档这三种相关实体发现方法的融合，有效地获取与查询实体相关的实体集合，从而避免了通过遍历的方式计算知识库中所有实体与查询实体的相关度来进行召回，极大地降低了计算量. 另一方面，通过在相关实体排序阶段引入多种不同类型的特征并基于这些特征学习排序模型，较为有效地处理新查询与新实体的推荐问题，从而缓解了传统的基于协同过滤的方法[25]由于依赖用户过往搜索行为数据所要面临的数据稀疏与冷启动问题. 其次，用户在搜索引擎中输入的查询的类别是开放领域的，如果采用先对查询进行分类，再基于查询类别进行实体推荐的方法，则不可避免地会受到分类错误传递的影响. 为解决这一挑战，Huang 等人[6]与 Fernandez-Tobias 等人[25]提出了领域无关的实体推荐方法，从而有效地解决了由于依赖领域相关特征所导致的只能先明确查询类别才能进行实体推荐[27-28]的问题. 最后，实体推荐任务旨在为用户输入的查询推荐出相关实体，从而帮助用户发现感兴趣的实体，提升用户的搜索体验. 如果只基于查询进行相关实体推荐，而不考虑用户偏好，则会导致具有不同实体偏好的用户在输入相同查询时获得的实体推荐结果完全一致，从而无法为这些用户提供更符合个人偏好的个性化推荐结果. 因此，为了更好地满足用户的信息发现需求，在实体推荐中考虑用户偏好因素至关重要. 然而，在搜索引擎中，无法显式获取用户对实体的偏好信息. 为解决这一挑战，之前的研究工作提出了两种思路：使用短期搜索历史[25-26]或使用长期搜索历史[6,27-28]建模用户对不同实体的偏好，从而为用户提供更满意的实体推荐结果. 其中，短期搜索历史由用户在当前搜索会话中输入的所有前序查询及其点击信息构成，隐含着用户的短期兴趣，有助于对用户在当前搜索会话中对不同实体的偏好进行建模. 而长期搜索历史则由用户在一个较长时间段内的所有搜索会话中的查询及其点击信息累积而成，隐含着用户的长期兴趣，有助于建模用户对不同实体的固有偏好.

近年来，虽然实体推荐方向上取得了较好的研究成果，但该研究方向仍然存在以下 5 个尚待解决的重要问题：

(1) 近年来，为了使推荐出的实体能与用户的偏好更相关，实体推荐已逐渐从非个性化[5,25-26]改进到个性化[6,27-28]. 此外，为了使推荐出的实体能与用户的搜索需求更相关，实体推荐也逐渐从上下文无关[5-6,27-28]改进到上下文相关[25-26]. 然而，要将用户在单个搜索会话中对实体的短期兴趣与用户在一段历史时期中对实体的长期兴趣进行有效建模与融合存在较大挑战，因此目前尚无研究工作能够同时为用户提供个性化且上下文相关的实体推荐结果. 为了提供能让用户更满意的实体推荐结果，需要针对上述挑战找到对应的解决方法.

(2) 由于实体指称类的查询(例如“奥巴马”)的搜索意图较易确定，因此目前的实体推荐方法[5-6,25-28]都只处理实体指称类的简单查询. 相比之

① https://en.wikipedia.org/wiki/A/B_testing

下,包含一个或多个实体指称(例如“奥巴马的教育履历”)以及未包含任何实体指称(例如“什么东西适合天冷时吃”)的复杂查询的搜索意图更难确定,目前的实体推荐方法均未对这两类查询做出处理.由于实体指称类查询在搜索引擎整体查询中占比有限,例如Li等人[92]的统计结果显示仅有48.8%的查询为实体指称类查询.因此,为了提升实体推荐结果的覆盖率,目前的主流商业搜索引擎,例如百度、Google均支持为复杂查询进行实体推荐,典型的查询示例为“奥巴马的教育履历”(百度)、“什么东西适合天冷时吃”(百度)、“Einstein education”(Google).为了处理复杂查询,需要对其背后的搜索意图进行语义理解,从而提供与之相关的实体推荐结果.这也是将实体推荐方法实际应用到大规模搜索引擎所要面临的挑战之一,因此也需要为其找到对应解决方法.

(3)在建模用户对实体的兴趣度时,目前大部分实体推荐方法[5,25-28]只使用了从知识库中获得的实体唯一标识、实体名称、实体类别以及实体属性等信息,只有很少的工作[6]使用了实体描述信息,而未有任何工作采用实体图片信息.实体描述信息对于实体兴趣度的建模具有重要作用.一方面,由于知识库对新出现实体的覆盖率往往不足,因此采用实体描述信息①有助于缓解新实体的覆盖率问题.另一方面,与实体名称相比,实体描述含有丰富的语义信息,有助于更好地对实体进行表示学习[93].实体图片也是实体推荐结果的重要组成部分,因此对于建模用户对实体的兴趣度同样具有重要作用.由于文本与图片属于不同模态的信息,因此如何利用多模态实体信息进行实体兴趣度建模,也是实体推荐方法需要解决的挑战之一.

(4)目前的大部分实体推荐方法[6,25-28]均依赖于大规模搜索日志中的用户实体点击数据进行模型训练,因此这些方法均不适用于解决系统冷启动时,在缺乏这些用户数据的情况下如何有效构建实体推荐系统的问题.虽然Blanco等人[5]采用人工标注的方式获得数据集,有助于缓解依赖用户行为数据的问题,但该方法标注的数据量小且标注出的相关性与真实用户的搜索需求之间不可避免地存在偏差.此外,上述所有方法均未对缺乏足够用户行为数据的冷门查询以及新实体的推荐问题进行专门研究与分析.冷启动问题也是将实体推荐方法应用到大规模搜索引擎所要面临的挑战之一,因此也需要为其找到对应解决方法.

(5)实体推荐系统目前主要应用于大规模搜索引擎,因此依赖于大规模知识库与海量真实用户数据进行模型训练与效果评测.大规模知识库的构建是一项极其耗时耗力的工作,因此目前的一部分工作[27-28]选择使用开放知识库,例如Freebase[9]、DBpedia[10]、YAGO[37]以及NELL[94],而一部分工作[5-6]则选择使用所在公司构建的专有知识库.由于缺少面向实体推荐任务的大规模开放用户搜索日志数据集,而人工标注数据集的方式往往存在以下两方面的缺点:①人工标注成本高昂且标注的数据量也往往有限;②人工标注的相关性与基于真实搜索用户行为所计算出的相关性之间不可避免地存在偏差.因此,目前的大部分实体推荐方法[6,25-28]均基于搜索日志与实体点击日志,自动从中生成模型训练所需的数据集.只有少部分方法[5]采用人工标注的方式获得数据集.由于大规模开放数据集对于实体推荐研究至关重要,因此缺少开放数据集也是该研究领域存在的问题之一.

4 推荐理由生成

在传统的推荐系统中,研究结果表明对推荐结果进行恰当且合理的解释有助于提升用户在透明度、可信度、有效性、接受度与满意度等方面的体验[95-102].虽然解释对于提升推荐系统的用户体验具有至关重要的作用,但很难对什么是好的解释进行统一定义,因为这往往取决于设计推荐系统时希望达到的目标.表2列出了对推荐结果进行解释的7大可能目标[103].不同目标之间可能是互补的,也可能是对立的,因此不可能兼具所有目标.例如,有效性能提升可信度,但说服性可能会降低有效性.在实体推荐系统中,用户希望能够迅速地理解推荐结果与其搜索需求间的相关性.因此,解释的目标重在帮助用户更快、更好地做出决策,使系统更易于使用,即侧重于有效性、效率、可信度以及满意度这4个目标.

表2 推荐系统可解释性的目标[103]

目标	定义
透明度(transparency)	解释推荐系统如何工作
可检视(scrutability)	让用户发现推荐是否准确
可信度(trust)	提升用户对推荐系统的信任
有效性(effectiveness)	帮助用户做出更好的决策
说服性(persuasiveness)	说服用户进行尝试或购买
效率(efficiency)	帮助用户更快地做出决策
满意度(satisfaction)	提升易用性或愉悦度

① 新实体的描述信息一般从实体对应的在线百科文章中获取.

具体地，在实体推荐中，如果系统只根据用户输入的查询返回相关实体推荐结果，而不对推荐结果进行必要的解释，用户可能不易理解为什么这些实体会被推荐给自己，进而会对推荐结果产生疑惑. 因此，实体推荐系统不仅需要准确地为用户提供与其信息需求相关的实体，还需要提供恰当且合理的推荐理由，以便于用户能够迅速地理解、相信并接受所推荐的实体结果. 以图 11 为例，当用户输入查询“俊介”后，如果在实体推荐结果中只展示实体名称，而不展示实体图片与推荐理由，用户可能较难理解这些实体与查询之间的相互关系. 从图中可以看出，推荐理由有两种不同的维度：集合推荐理由与实体推荐理由. 集合推荐理由“与俊介一样娇小可爱的萌宠”与“那些家喻户晓的明星宠物”描述了各自实体集合中的 4 个实体与查询“俊介”的共同特征，让用户能够迅速地理解为什么会推荐这些实体. 而每个实体下方所展示的实体推荐理由，则能帮助用户理解单个实体与其搜索需求之间存在怎样的相关性.

图 11　带有推荐理由的实体推荐结果示例

为实体推荐结果增加推荐理由存在诸多挑战. 首先，由于实体推荐结果是按照多实体聚合与单实体排列相结合的方式进行呈现，这就要求在实体集合与单个实体这两种不同维度上都要进行必要的解释. 其次，用户在搜索引擎中输入的查询实体的规模往往非常庞大，而实体也丰富多变，这为集合推荐理由与实体推荐理由的生成带来了极大的挑战. 最后，如果要在实体下方有限的空间中让用户不经过任何点击或跳转就能直接看到实体推荐理由的全部内容，则需要对不符合长度要求的实体推荐理由进行压缩①. 图 12 概括了实体推荐可解释性研究需要解决的主要任务.

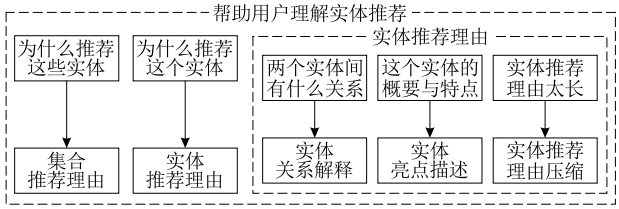

图 12　实体推荐可解释性研究的主要任务

4.1　集合推荐理由

集合推荐理由旨在解释被推荐实体集合与用户查询所蕴含的搜索需求间存在怎样的相关性，从而方便用户判断是否有兴趣进一步了解其中包含的推荐实体. 此外，为了让用户感知到更好且更直接的相关性，在集合推荐理由中提供与用户搜索需求相关的个性化解释信息至关重要. 例如，与“相关动物”相比，“与俊介一样娇小可爱的萌宠”更能提升用户对实体推荐结果的接受度与信任度，因为后者不仅反映了查询“俊介”与被推荐实体集合之间的共同特点，还预测并概括了用户在搜索该查询时的潜在搜索需求之一. 因此，按照生成集合推荐理由时是否考虑用户输入的查询，可以将生成方法分为两类：与查询无关的方法以及与查询相关的方法.

与查询无关的集合推荐理由生成方法大致可以分为基于标签与基于模板这两类方法：

(1) 在当前实体推荐系统中，集合推荐理由通常根据当前实体集合中所有被推荐实体的粗粒度分类标签进行描述[104]，例如“相关动物”、“相关植物”等. 但这种集合推荐理由粒度过粗且形态单一，难以让用户一目了然地理解推荐缘由，因此在提升用户对实体推荐结果的接受度与信任度上的帮助有限. 为缓解这一问题，可以通过将粗粒度分类标签优化为细粒度分类标签[105-108]. 例如，将“相关动物”细化为“相关宠物犬”，“相关植物”细化为“相关多肉植物”，从而能够向用户传达更具体且信息更丰富的推荐缘由.

(2) 在一些搜索引擎的实体推荐结果中，也会使用“用户还搜索了”或“其他人还搜”等基于搜索模式的固定语句作为集合推荐理由. 这种固定模板式的推荐理由在电子商务网站及社交网站的推荐系统中也经常被采用. 例如，大部分电子商务网站的推荐系统中常采用“购买此商品的顾客也同时购买”、“浏览此商品的顾客也同时浏览”等基于购买模式的推

① 根据实体推荐结果的产品设计，要求将实体推荐理由全部展现在实体下方的有限空间中，因此必须将其长度压缩到固定字数以内. 而集合推荐理由的展现空间一般较为充裕，通常不需要进行压缩.

荐理由[99].而基于社交网络的内容推荐系统中也常采用“你的朋友也在看”、“你的好友也关注了”等基于社交关系的推荐理由[109].

这两种方法能够在一定程度上解决使用实体粗粒度分类标签作为集合推荐理由带来的用户接受度与信任度不高的问题.但不足之处在于这些推荐理由仍过于简化且形态单一,无法为用户提供能够反映实体集合与查询所蕴含的搜索需求间存在怎样的相关性的详细解释,因此也很难显著提升用户的接受度与信任度.

与查询相关的集合推荐理由生成方向上的研究工作不多.如果将问题建模为寻找查询实体与实体集合所包含的各实体间存在的共同特点作为推荐理由,则可以采用基于知识图谱的方法[110-111],从中获取多个实体的共同属性描述(例如“长毛犬”)后再基于模板(例如“与[实体]类似的[属性描述]”)填充对应词槽生成解释文本.此外,也可以采用基于情感词短语生成解释文本的思路[112-113],从评论文本中抽取出查询实体与实体集合所包含的各实体共同具有的情感词(例如“娇小可爱”),然后再基于模板(例如“与[实体]一样[情感词]的[细粒度分类标签]”)填充情感词等词槽生成解释文本.这两种方法都可以处理与查询相关的集合推荐理由生成问题,但不足之处在于覆盖率有限且解释模板多样性较低.

4.2 实体推荐理由

为了帮助用户更好地理解推荐出的实体,需要为实体生成恰当且合理的推荐理由.此外,为了能够让用户高效地获知每个实体被推荐的理由,需要在实体下方有限的空间中直接展示出实体推荐理由的全部内容.因此,在为被推荐实体生成实体推荐理由之后,还需要进一步对生成的实体推荐理由进行压缩,以使其符合搜索引擎要求的字数限制.下面分别对实体推荐理由生成与压缩这两个子任务进行介绍与分析.

4.2.1 实体推荐理由生成

实体推荐理由旨在解释被推荐实体与用户查询所蕴含的搜索需求间存在怎样的相关性,从而方便用户判断是否有兴趣进一步了解该实体.

从图1中显示的查询“奥巴马”的实体推荐结果示例中可以看出,其中部分实体的推荐理由与查询相关,而部分与查询无关.例如,在实体“米歇尔·奥巴马”下方所展示的推荐理由“92年结婚并育有俩女儿”,提供了有关查询实体“奥巴马”与被推荐实体“米歇尔·奥巴马”这两个实体之间关系的解释.在被推荐实体“唐纳德·特朗普”下方所展示的推荐理由“美国总统奥巴马继任者”也解释了这两个实体间的关系.而在被推荐实体“威廉·杰斐逊·克林顿”与“马丁·路德·金”下方所展示的推荐理由“第42任美国总统”与“美国黑人民权运动领袖”,则只与当前实体有关.

在实体推荐结果下方展现的推荐理由,能够帮助用户快速了解查询实体与被推荐实体间的关系,或者被推荐实体的关键特征和信息,从而帮助用户理清这些实体与自己所输入的查询之间存在的关系或联系,因此有助于提升实体推荐结果的易理解性[31,33-34,114].按照生成实体推荐理由时是否考虑用户输入的查询,可以将生成方法分为两类:与查询相关的方法以及与查询无关的方法.

(1)与查询相关的实体推荐理由生成

生成与查询相关的实体推荐理由,目前的方法主要侧重于解决如下任务:给定两个实体及其关系构成的三元组$\langle e_i, r_k, e_j \rangle$,生成能够解释$e_i$与$e_j$间给定关系$r_k$的自然语言句子①.现有解决方法大致可以分为以下两类:

①基于模板的方法.模板可以由人工标注或自动学习获得.采用基于人工标注模板的方法生成实体关系解释句子[29]简单易行,但却存在两个方面的局限.首先,这种方法需要为每一种实体关系人工标注一定数量的模板,由于关系众多且人工标注成本高昂,导致该方法很难应用于大规模实体关系解释句子生成任务上.其次,虽然该方法能达到很高的准确率,但由于人工标注的模板数量往往有限,因此召回率常常较低.

为了解决上述问题,Voskarides等人[30]提出基于知识图谱自动获得特定实体关系r_k的解释句子模板,然后在为具备同样关系的新三元组$\langle e_h, r_k, e_t \rangle$生成关系解释句子时,只需要在模板中将新实体对$e_h$与$e_t$及其属性填入对应的槽进行实例化即可.这种方法能够有效地处理高频实体关系的解释,例如,基于大量关系为“演员-主演-电影”的三元组及其关系解释句子构建出实体依赖关系图,然后再根据该关系图自动学习出模板(例如“[电影]是[公司]出品的[电影类型]电影,由[演员]领衔主演.”).当给定新三元组〈尼尔·塞西,主演,奇幻森林〉时,基于该模板与知识图谱中的实体属性信息(例如电影奇幻森林的出品公司与电影类型),即可生成关系解释句子“奇幻森林是迪士尼出品的奇幻真人动画电影,由

① 虽然可以直接用知识图谱中已存在的实体关系或通过实体关系预测[143-146]得到的关系,例如“配偶”与“子女”,来对实体间关系进行注解,但这些关系类型的解释性与信息量往往不足,无法用于前述实体推荐结果中对实体关系进行详细刻画或解释.为了更好地解释两个实体间的给定关系,提供有关该关系的详细解释或佐证句子至关重要.

尼尔·塞西领衔主演.”.但该方法的不足之处在于知识图谱中实体关系与实体属性的覆盖率往往有限,会导致在实际大规模实体推荐系统应用中的召回率较低.此外,由于生成的解释性句子的表达方式有限且固定,会导致实体推荐理由的多样性较低.

② 基于句子检索的方法. Voskarides 等人[32]首先对获取实体关系解释句子这一任务进行了研究,并提出了一种基于单文档(pointwise)排序模型的实体关系解释句子获取方法:首先从给定文档中抽取出候选句子,然后再基于人工设计的特征对这些候选句子进行排序,从而获得与给定三元组所对应的解释句子.虽然该方法在小规模数据上取得了较好的实验结果,但在应用于大规模、真实任务时存在两方面的缺点.首先,大规模训练数据对基于有监督机器学习方法的排序模型而言至关重要.由于人工标注成本高昂,采用这种方法构建大规模训练数据太过昂贵.其次,该方法使用人工设计的特征训练排序模型.由于在特征抽取过程中不可避免地会出现错误,因此基于人工特征的排序模型的效果不可避免地会受到错误传递的影响.

为解决上述问题,Huang 等人[31]提出了一种基于文档对(pairwise)排序模型获取实体关系解释句子的方法,并使用 CNN 自动从大量训练样本中学习出相关特征,从而无需依靠人工设计特征.此外,该工作还提出了一种借助于搜索引擎点击日志自动构建大规模训练数据的方法,从而无需依靠人工标注获取训练数据.图 13 显示了该模型的网络结构.该网络的输入为一个由查询 $q_s=\langle e_h, r_k, e_t\rangle$ 及一对网页标题 t_i 与 t_j 所构成的三元组 $\langle q_s, t_i, t_j\rangle$.然后,采用带有相同结构及参数的 CNN 网络将 q_s、t_i 及 t_j 转化为各自对应的向量表示 $v(q_s)$、$v(t_i)$ 及 $v(t_j)$.最后,该模型的学习目标是优化表示函数 $v(\cdot)$,使得与 q_s 更相关的 t_i 能得到更高的相似度得分,即

$$cos(v(q_s), v(t_i)) > cos((v(q_s), v(t_j)),$$
$$\forall q_s, t_i, t_j \text{ 给定 } rel(q_s, t_i) > rel(q_s, t_j) \quad (11)$$

实验结果表明该方法显著优于多个稳健的基线方法.这种方法能够基于海量互联网网页获取大规模且多样性较高的实体关系解释句子,但不足之处在于准确率较低,很难直接应用到实际推荐系统中.

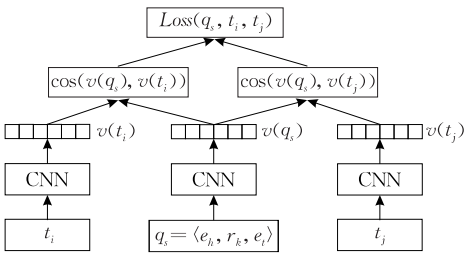

图 13　Huang 等人[31]提出的文档对排序模型的网络结构

(2) 与查询无关的实体推荐理由生成

生成与查询无关的实体推荐理由,目前的方法主要侧重于解决如下任务:给定一个实体 e 及其描述句子 $sent$,生成能够描述该实体独特之处的简短、精炼的自然语言表述 eh(即“实体亮点”).以实体“贝拉克·奥巴马”(e)为例,给定一个描述该实体信息的句子“贝拉克·奥巴马是美国政治人物,从 2009 年至 2017 年任第 44 任美国总统.”($sent$),实体亮点生成任务的目标是从 $sent$ 中生成出一个与实体 e 相关的自然语言表述“第 44 任美国总统”(eh).

与查询无关的实体推荐理由生成任务,目前以基于文本生成的方法为主.为解决这一问题,Huang 等人[33]提出了一种基于序列到序列学习(Sequence-to-Sequence Learning,简称 Seq2Seq)[115]的实体亮点生成模型.图 14 显示了引入注意力机制[116-118]、

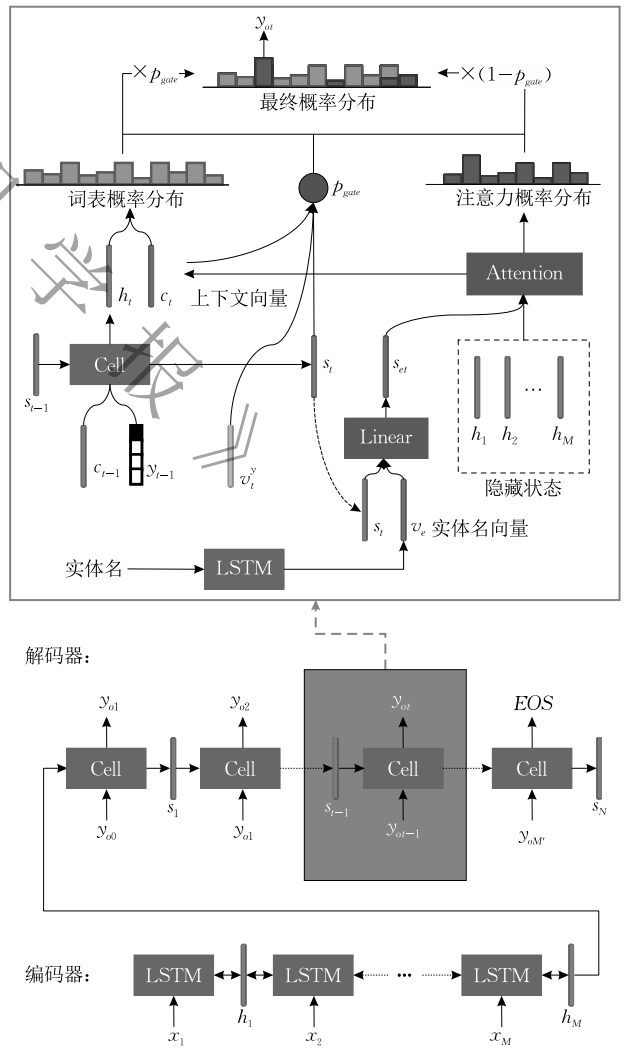

图 14　Huang 等人[33]提出的基于 Seq2Seq 的实体亮点生成模型的网络结构

复制机制[119-123]以及实体辅助信息的 Seq2Seq 模型的网络结构.与普通 Seq2Seq 模型相比,该生成模型在解码阶段引入了多种不同的机制,以生成质量更高且与实体更相关的亮点.首先,通过引入注意力机制,使解码器在生成当前词时能够区分源句子中各个词的重要性.其次,通过引入复制机制,使解码器能够处理稀有词的生成问题,从而生成出源文本中的重要的低频词.然后,通过将实体名作为辅助信息引入到解码过程中,可以指引模型生成与给定实体更相关的亮点.最后,该生成模型还进一步引入了覆盖机制[124-125],以缓解生成重复词的问题.虽然该方法取得了令人满意的结果,但不足之处在于生成实体亮点时,并未考虑用户输入的查询.因此,相同实体在不同查询下所展现的实体亮点均固定不变,可能出现推荐理由与用户搜索需求弱相关或不相关的情况.

4.2.2 实体推荐理由压缩

从图 11 中可以看出,如果要将实体推荐理由全部直接展现在实体下方的有限空间中,则必须将其长度压缩到固定字数以内,同时仍然是一个流畅的自然语言表述.此外,为了让用户能够快速了解实体与搜索需求间存在的相关性或实体的关键特征,压缩后的实体推荐理由必须保持足够的信息量与吸引力.例如,将较长的推荐理由"因酷像美国总统奥巴马而在世界各地以模仿奥巴马来赚钱养家"和"与奥巴马相貌几乎一模一样"分别压缩并改写成"长相酷似奥巴马"和"与奥巴马长相神似".

与实体推荐理由压缩任务较为接近的是句子压缩任务.目前的句子压缩方法大致可以分为以下三类.第一类方法主要基于删除源句子中的某些词或成分的方法来对源句子进行压缩[126-132].第二类方法则加入了除删除外的更多操作,通过对源句子中的词进行调序、替换、插入以及删除操作,实现对源句子进行压缩[133-134].第三类方法主要基于神经机器翻译模型,采用"翻译"的思路将源句子进行压缩[33,135].与句子压缩相比,实体推荐理由压缩不仅需要对源推荐理由中冗余的词进行删除并对其中较长的词或短语进行替换,还需要提升信息量与吸引力,并生成出简短、精炼且符合特定长度要求的自然语言表述.因此,上述句子压缩方法均无法直接用于实体推荐理由压缩任务.

为解决这一问题,Huang 等人[34]提出了一种基于统计机器翻译的实体推荐理由压缩模型,从而将较长的推荐理由压缩并生成出符合特定长度要求、流畅、有信息量且对用户有吸引力的短推荐理由.该方法利用搜索引擎点击日志中的"网页标题-查询"对构建所需的单语平行语料,并利用短且能吸引用户浏览的新闻标题训练吸引力模型.该模型主要由翻译模型、语言模型、长度模型以及吸引力模型构成:

$$\begin{aligned}P(se|le)=&\lambda_{tm}logP_{tm}(\overline{le}_1^l,\overline{se}_1^l)+\lambda_{lm}logP_{lm}(se)+\\&\lambda_{lf}logP_{lf}(se)+\lambda_{hl}logP_{hl}(se)+\\&\lambda_{ss}logP_{ss}(se)\end{aligned}\tag{12}$$

在上述公式中,le 与 se 分别表示长推荐理由与经过压缩后的推荐理由.$P_{tm}(\overline{le}_1^l,\overline{se}_1^l)$为翻译模型,$P_{lm}(se)$为语言模型.$P_{lf}(se)$为专门设计的长度惩罚函数,用于促使模型生成尽可能简短且精炼的推荐理由.$P_{hl}(se)$为在新闻标题语料上训练得到的语言模型,用于促使模型生成的推荐理由在词汇和表达方式上与新闻标题尽可能接近.$P_{ss}(se)$为句子结构相似度函数,用于促使模型生成与人工编写的推荐理由在句法风格上尽可能接近的推荐理由.实验结果表明该方法可以将任意类型的长推荐理由压缩成质量较好的短推荐理由.实验结果还显示,采用新闻标题这种短且能吸引用户眼球的语料训练吸引力模型,能够显著提升实体推荐理由压缩的效果.相似的结论在 Klerke 等人[136]的工作中也得到了进一步验证:使用吸引用户眼球的语料能够显著提升句子压缩的效果.

4.3 推荐理由评价指标

推荐理由生成以文本生成方法为主,因此评价指标相应地以 BLEU[137]、ROUGE[138]以及人工评价这三种方式为主[33-34].BLEU 与 ROUGE 分别是机器翻译与文本摘要中广泛采用的评价指标.BLEU 用于衡量模型产出的翻译结果与标准参照译文间的相似度.而 ROUGE 则主要用于衡量模型产出的摘要与人撰写的标准参照摘要间的相似度.人工评价与机器翻译中所采用的人工评价方法[139]类似,主要基于流畅度与可用度这两个子指标对模型生成的推荐理由的质量进行评价.

4.4 推荐理由生成方法总结

为便于用户能够迅速地理解、相信并接受实体推荐结果,还需要为被推荐的实体集合以及每一个被推荐实体生成恰当且合理的推荐理由.其中集合推荐理由旨在解释为什么要将给定实体集合推荐给用户,而实体推荐理由则侧重于解释为什么要将给定实体推荐给用户.具体地,给定查询实体 e_q 与被推荐实体集合 $E=\{e_1,e_2,\cdots,e_n\}$,集合推荐理由生成任务需要解决的问题可以表示为 $G_E(e_q,E)\mapsto S_{qE}$,

而实体推荐理由生成任务需要解决的问题则为 $G_e(e_q,e_i)\mapsto S_{qe}$. 集合推荐理由生成任务的难点在于如何生成能够描述一组实体共同特点的句子 S_{qE}，而实体推荐理由生成任务的难点则在于如何生成能够解释两个实体 e_q 与 e_i 间关系的句子 S_{qe}. 虽然这两个任务的目标不同，但都可以采用基于模板的方法[29-30,109-111]. 然而，由于很难直接从网页中获得能够对给定实体集合中的所有实体的共同特点进行描述的句子，因此在实体推荐理由生成任务中所主要采用的两类方法：基于句子检索的方法[31-32]与基于句子的文本生成方法[33]，很难直接应用于集合推荐理由生成任务. 在效果上，基于模板的方法能够生成准确率较高的推荐理由，但却存在以下不足之处. 首先，模板覆盖率往往有限且模板多样性较低. 此外，由于这类方法通常依赖于知识图谱中的实体分类及属性等信息，而知识图谱往往缺乏对新实体的覆盖，因此一旦存在新实体时，这类方法就会生成失败. 虽然基于句子检索的方法与基于句子的文本生成方法都能够在一定程度上缓解上述问题，但获得的推荐理由准确率往往较低，很难直接应用于实际推荐系统中. 由于上述方法的准确率与覆盖率无法同时达到理想效果，因此很难将其实际应用到大规模搜索引擎中，这也是推荐理由生成任务所面临的主要挑战，因此需要为其找到对应解决方法.

5　总结及未来研究方向

实体推荐旨在帮助用户探索并发现与其搜索需求相关的实体，已成为现代搜索引擎必不可少的功能之一. 从 2013 年开始，面向搜索引擎的实体推荐得到了研究人员的广泛关注，相关研究成果陆续被发表出来[5-6,25-28]. 由于这一方向的研究时间相对较短，虽然目前的工作已取得了较好的研究成果，但仍然存在一些值得深入探索的问题，以下几点有可能成为未来研究方向：

(1) 基于历史搜索查询及点击信息的实体链接：当前实体推荐系统所处理的搜索查询只包含一个实体指称而没有其他任何上下文信息，单纯从这个实体指称本身无法推断用户究竟想搜索的是哪一个实体. 而用户在同一个搜索会话中的搜索查询之间具有一定的任务相关性，因此可以借助于本次会话中的历史查询及其点击信息对当前查询进行实体链接. 历史查询中的实体指称本身也是有歧义的，因此可以采用联合的方法同时对历史查询中的实体指称和当前查询的实体指称进行实体链接.

(2) 为更复杂的搜索查询推荐相关实体：目前搜索引擎的实体推荐系统处理的搜索查询类型相对简单，大部分是只包含一个实体指称的查询(例如“奥巴马”). 这一类查询的搜索意图较为明确，即与该实体指称相关的信息. 对于实体指称具有描述信息或者包含多个实体指称的复杂查询，当前的实体推荐系统大部分不作处理，因为这一类查询的主要搜索需求更难确定. 然而与只有一个实体指称的查询相比，这一类搜索查询的优点是包含了更丰富的上下文信息，有助于更好地理解查询. 例如，可以利用查询中的上下文信息对实体指称进行更加精确地实体链接. 在未来面向搜索引擎的实体推荐系统中，可以考虑处理更复杂的搜索查询. 例如，可以通过对搜索查询进行深度语义理解后，获得能够代表用户信息需求的核心实体，再基于该实体进行实体推荐.

(3) 在实体推荐模型中引入更多维度的用户偏好信息：借助于用户偏好信息，可以更好地理解不同用户对不同实体的偏好程度，从而能够推荐出用户更喜欢的实体. 但目前的实体推荐方法在建模用户偏好时，只考虑了用户对实体的历史点击率[6,27-28]，而未考虑用户历史上搜索过的查询、点击与浏览过的网页文档. 用户的搜索、点击与浏览行为都能反映用户对不同信息的偏好，对更准确地建模用户的信息需求具有重要作用. 例如，Wu 等人[140]的研究表明，在查询建议(query suggestion)任务中引入用户的网页点击与浏览信息，有助于为用户提供更准确且多样化的结果. Ahmad 等人[141]的研究工作也表明，采用基于多任务学习的方式同时训练网页排序与查询建议两个任务，对各自效果均有提升. 这表明在个性化实体推荐模型的未来研究中，可以考虑引入更多维度的用户偏好信息.

(4) 在实体推荐模型中引入更丰富的上下文信息. 同一搜索会话中的上下文信息有助于更准确地理解用户输入的当前查询背后的信息需求，因此，在实体推荐模型中引入上下文信息能够有效提升推荐效果[26]. 但目前的方法要么只考虑了用户在同一搜索会话中输入的前序查询序列[26]，要么只考虑了用户在搜索会话中点击过的实体[25]. 为了推荐出与用户当前信息需求更相关的实体，在上下文相关的实体推荐模型中，将前序查询及其对应点击信息(对实体或网页文档的点击)都考虑进来，是未来值得研究的方向.

(5) 提高实体推荐的可解释性. 为实体推荐结

果提供恰当且合理的推荐理由，能够帮助用户迅速地理解、相信并接受推荐结果.但目前的方法只解决了其中一部分问题，该方向的工作至少存在以下待解决或待改进的问题.首先，实体关系可能会随时间发生变化，因此需要对实体间不断变化的关系进行解释.例如，可以基于时间轴的方式对实体关系进行解释[29].其次，目前的方法只侧重于为存在直接关系的实体对生成关系解释句子，无法很好地处理存在非直接关系的实体对.最后，为进一步提升相关性，在展示推荐理由时，需要考虑为不同查询选择最合适的推荐理由.

(6)基于神经网络的多模态实体推荐模型.从图 1 与图 11 中可以看出，实体图片也是实体推荐结果的重要组成部分.如果一个实体的图片能够非常生动地刻画该实体的最典型特征(例如图 11 中长相奇特的“不爽狗”的图片)，就相当于从视觉上对实体的亮点进行了呈现.而用户一旦对一个实体的图片产生了兴趣，自然地就会想进一步了解该实体.因此，实体图片在一定程度上也起到了推荐理由的作用，是对文本推荐理由的有效补充，这也契合了常说的“一图胜千言”.这说明实体图片在实体推荐系统中也具有重要作用.目前虽然已有研究工作探索将电影封面图作为个性化因素引入到电影推荐系统中[142]，但目前的实体推荐研究工作均未考虑实体图片，也尚无工作对实体图片在实体推荐系统中的作用与效果进行研究与分析.由于神经网络能够在一个统一的空间中对不同模态(例如文本、图像、语音等)的数据进行表示，因此采用神经网络构建基于多模态实体信息(例如实体名称、实体描述文本、实体图片等)的端到端的实体推荐系统，也是未来值得探索的研究方向.

参 考 文 献

[1] Chilton L B, Teevan J. Addressing people's information needs directly in a Web search result page//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 27-36

[2] Bernstein M S, Teevan J, Dumais S, et al. Direct answers for search queries in the long tail//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Austin, USA, 2012: 237-246

[3] Cao Huanhuan, Jiang Daxin, Pei Jian, et al. Context-aware query suggestion by mining click-through and session data//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 875-883

[4] Mei Qiaozhu, Zhou Dengyong, Church K. Query suggestion using hitting time//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa, USA, 2008: 469-478

[5] Blanco R, Cambazoglu B B, Mika P, Torzec N. Entity recommendations in Web search//Proceedings of the 12th International Semantic Web Conference. Sydney, Australia, 2013: 33-48

[6] Huang Jizhou, Ding Shiqiang, Wang Haifeng, Liu Ting. Learning to recommend related entities with serendipity for Web search users. ACM Transactions on Asian and Low-Resource Language Information Processing, 2018, 17(3): 25:1-25:22

[7] Guo Jiafeng, Xu Gu, Cheng Xueqi, Li Hang. Named entity recognition in query//Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Boston, USA, 2009: 267-274

[8] Pound J, Mika P, Zaragoza H. Ad-hoc object retrieval in the Web of data//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 771-780

[9] Bollacker K, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge//Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. Las Vegas, USA, 2008: 1247-1250

[10] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a Web of open data//Proceedings of the 6th International the Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference. Busan, Korea, 2007: 722-735

[11] Hachey B, Radford W, Nothman J, et al. Evaluating entity linking with Wikipedia. Artificial Intelligence, 2013, 194(1): 130-150

[12] Blanco R, Ottaviano G, Meij E. Fast and space-efficient entity linking for queries//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. Shanghai, China, 2015: 179-188

[13] McNamee P, Dang H T. Overview of the TAC 2009 knowledge base population track//Proceedings of the Text Analysis Conference (TAC). Gaithersburg, USA, 2009: 111-113

[14] Ji H, Grishman R, Dang H T, et al. Overview of the TAC 2010 knowledge base population track//Proceedings of the Text Analysis Conference (TAC). Gaithersburg, USA, 2010: 1-25

[15] Carmel D, Chang M-W, Gabrilovich E, et al. ERD'14: Entity recognition and disambiguation challenge. SIGIR Forum, 2014, 48(2): 63-77

[16] Feng Yansong, Han Zhe, Zhang Kun. Overview of the NLPCC 2015 shared task: Entity recognition and linking in search queries//Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing. Nanchang, China, 2015: 550-556

[17] Liu Xiaohua, Li Yitong, Wu Haocheng, et al. Entity linking for tweets//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 1304-1311

[18] Guo S, Chang M-W, Kiciman E. To link or not to link? A study on end-to-end tweet entity linking//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, USA, 2013: 1020-1030

[19] Hasibi F, Balog K, Bratsberg S E. Entity linking in queries: Tasks and evaluation//Proceedings of the 2015 International Conference on the Theory of Information Retrieval. Northampton, USA, 2015: 171-180

[20] Ji Heng, Nothman J, Hachey B. Overview of TAC-KBP2014 entity discovery and linking tasks//Proceedings of the Text Analysis Conference (TAC). Gaithersburg, USA, 2014: 1333-1339

[21] Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 160-167

[22] Ratinov L, Roth D, Downey D, Anderson M. Local and global algorithms for disambiguation to Wikipedia//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, USA, 2011: 1375-1384

[23] He Zhengyan, Liu Shujie, Li Mu, et al. Learning entity representation for entity disambiguation//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 30-34

[24] Yamada I, Shindo H, Takeda H, Takefuji Y. Joint learning of the embedding of words and entities for named entity disambiguation//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany, 2016: 250-259

[25] Fernandez-Tobias I, Blanco R. Memory-based recommendations of entities for Web search users//Proceedings of the 25th ACM International Conference on Information and Knowledge Management. Indianapolis, USA, 2016: 35-44

[26] Huang Jizhou, Zhang Wei, Sun Yaming, et al. Improving entity recommendation with search log and multi-task learning //Proceedings of the 26th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 4107-4114

[27] Yu Xiao, Ma Hao, Hsu B J P, Han Jiawei. On building entity recommender systems using user click log and freebase knowledge//Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA, 2014: 263-272

[28] Bi Bin, Ma Hao, Hsu B J P, et al. Learning to recommend related entities to search users//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. Shanghai, China, 2015: 139-148

[29] Althoff T, Dong X L, Murphy K, et al. TimeMachine: Timeline generation for knowledge-base entities//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 19-28

[30] Voskarides N, Meij E, de Rijke M. Generating descriptions of entity relationships//Proceedings of the 39th European Conference on Information Retrieval. Aberdeen, Scotland, 2017: 317-330

[31] Huang Jizhou, Zhang Wei, Zhao Shiqi, et al. Learning to explain entity relationships by pairwise ranking with convolutional neural networks//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 4018-4025

[32] Voskarides N, Meij E, Tsagkias M, et al. Learning to explain entity relationships in knowledge graphs//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 564-574

[33] Huang Jizhou, Sun Yaming, Zhang Wei, et al. Entity highlight generation as statistical and neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1860-1872

[34] Huang Jizhou, Zhao Shiqi, Ding Shiqiang, et al. Generating recommendation evidence using translation model//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 2810-2816

[35] Milne D, Witten I H. Learning to link with Wikipedia//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa, USA, 2008: 509-518

[36] Olieman A, Azarbonyad H, Dehghani M, et al. Entity linking by focusing DBpedia candidate entities//Proceedings of the 21st International Workshop on Entity Recognition & Disambiguation. Gold Coast, Australia, 2014: 13-24

[37] Suchanek F M, Kasneci G, Weikum G. YAGO: A core of semantic knowledge//Proceedings of the 16th International Conference on World Wide Web. Banff, Canada, 2007: 697-706

[38] Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 782-792

[39] Gottipati S, Jiang J. Linking entities to a knowledge base with query expansion//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 804-813

[40] Cucerzan S. Large-scale named entity disambiguation based on Wikipedia data//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Prague, Czech Republic, 2007: 708-716

[41] Eiselt A, Figueroa A. A two-step named entity recognizer for open-domain search queries//Proceedings of the 6th International Joint Conference on Natural Language Processing. Nagoya, Japan, 2013: 829-833

[42] Jain A, Pennacchiotti M. Domain-independent entity extraction from Web search query logs//Proceedings of the 20th International Conference Companion on World Wide Web. Hyderabad, India, 2011: 63-64

[43] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. University of Michigan, USA, 2005: 363-370

[44] Han Xianpei, Zhao Jun. NLPR_KBP in TAC 2009 KBP Track: A two-stage method to entity linking//Proceedings of the Text Analysis Conference (TAC). Gaithersburg, USA, 2009: 1-8

[45] Zhang Wei, Sim Yan Chuan, Su Jian, Tan Chew Lim. Entity linking with effective acronym expansion, instance selection, and topic modeling//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011: 1909-1914

[46] Tan Chuanqi, Wei Furu, Ren Pengjie, et al. Entity linking for queries by searching Wikipedia sentences//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 68-77

[47] Guo Yuhang, Qin Bing, Li Yuqin, et al. Improving candidate generation for entity linking//Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems. Salford, UK, 2013: 225-236

[48] Dredze M, McNamee P, Rao D, et al. Entity disambiguation for knowledge base population//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China, 2010: 277-285

[49] Zheng Zhicheng, Li Fangtao, Huang Minlie, Zhu Xiaoyan. Learning to link entities with knowledge base//Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA, 2010: 483-491

[50] Francis-Landau M, Durrett G, Klein D. Capturing semantic similarity for entity linking with convolutional neural networks//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 1256-1261

[51] Sun Yaming, Lin Lei, Tang Duyu, et al. Modeling mention, context and entity with neural networks for entity disambiguation //Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 1333-1339

[52] Fang Wei, Zhang Jianwen, Wang Dilin, et al. Entity disambiguation by knowledge and text jointly embedding//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany, 2016: 260-269

[53] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality//Proceedings of the Neural Information Processing Systems Conference. Lake Tahoe, USA, 2013: 3111-3119

[54] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space//Proceedings of the International Conference on Learning Representations. Scottsdale, Arizona, 2013: 1-12

[55] Gupta N, Singh S, Roth D. Entity linking via joint encoding of types, descriptions, and context//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2681-2690

[56] Hasibi F, Balog K, Bratsberg S E. Entity linking in queries: Efficiency vs. effectiveness//Proceedings of the 39th European Conference on Information Retrieval. Aberdeen, Scotland, 2017: 40-53

[57] Mueller D, Durrett G. Effective use of context in noisy entity linking//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 1024-1029

[58] He Zhengyan, Liu Shujie, Song Yang, et al. Efficient collective entity linking with stacking//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 426-435

[59] Globerson A, Lazic N, Chakrabarti S, et al. Collective entity resolution with multi-focal attention//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 621-631

[60] Nie Feng, Zhou Shuyan, Liu Jing, et al. Aggregated semantic matching for short text entity linking//Proceedings of the 22nd Conference on Computational Natural Language Learning. Brussels, Belgium, 2018: 476-485

[61] Han Xianpei, Sun Le. A generative entity-mention model for linking entities with knowledge base//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, USA, 2011: 945-954

[62] Taylor Cassidy Z C, Artiles J, Ji H, et al. CUNY-UIUC-SRI TAC-KBP2011 entity linking system description//Proceedings of the Text Analysis Conference (TAC). Gaithersburg, USA, 2011: 1-13

[63] Graus D, Kenter T, Bron M, et al. Context-based entity linking-university of amsterdam at tac 2012//Proceedings of the Text Analysis Conference (TAC). Gaithersburg, USA, 2012: 1-6

[64] Yu Xiao, Ren Xiang, Sun Yizhou, et al. Personalized entity recommendation: A heterogeneous information network approach //Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA, 2014: 283-292

[65] Sun Zhu, Yang Jie, Zhang Jie, et al. Recurrent knowledge graph embedding for effective recommendation//Proceedings of the 12th ACM Conference on Recommender Systems. Vancouver, Canada, 2018: 297-305

[66] Bron M, Balog K, de Rijke M. Related entity finding based on co-occurrence//Proceedings of the Eighteenth Text REtrieval Conference. Gaithersburg, USA, 2009: 1-4

[67] Balog K, Serdyukov P, De Vries A P. Overview of the trec 2010 entity track//Proceedings of the 19th Text REtrieval Conference. Gaithersburg, Maryland, USA, 2010: 1-13

[68] Reinanda R, Meij E, Pantony J, et al. Related entity finding on highly-heterogeneous knowledge graphs//Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Barcelona, Spain, 2018: 330-334

[69] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3(4-5): 993-1022

[70] Yang Z, Nyberg E. Leveraging procedural knowledge for task-oriented search//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 513-522

[71] Ponzetto S P, Strube M. Knowledge derived from Wikipedia for computing semantic relatedness. Journal of Artificial Intelligence Research, 2007, 30(1): 181-212

[72] Liu J, Birnbaum L. Measuring semantic similarity between named entities by searching the Web directory//Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Fremont, USA, 2007: 461-465

[73] Tuarob S, Mitra P, Lee G C. Taxonomy-based query-dependent schemes for profile similarity measurement//Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search. Portland, USA, 2012: 8:1-8:6

[74] Ollivier Y, Senellart P. Finding related pages using green measures: An illustration with Wikipedia//Proceedings of the 22nd AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2007: 1427-1433

[75] Yeh E, Ramage D, Manning C D, et al. WikiWalk: Random walks on Wikipedia for semantic relatedness//Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing. Suntec, Singapore, 2009: 41-49

[76] Sun Yizhou, Han Jiawei, Yan Xifeng, et al. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003

[77] Yu Xiao, Sun Yizhou, Norick B, et al. User guided entity similarity search using meta-path selection in heterogeneous information networks//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 2025-2029

[78] Strube M, Ponzetto S P. WikiRelate! computing semantic relatedness using Wikipedia//Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference. Boston, USA, 2006: 1419-1424

[79] Milne D, Witten I H. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links//Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy. Chicago, USA, 2008: 25-30

[80] Hoffart J, Seufert S, Nguyen D B, et al. KORE: Keyphrase overlap relatedness for entity disambiguation//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 545-554

[81] Gabrilovich E, Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis//Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India, 2007: 1606-1611

[82] Aggarwal N, Buitelaar P. Wikipedia-based distributional semantics for entity relatedness//Proceedings of the 2014 AAAI Fall Symposium Series. Quebec, Canada, 2014: 2-9

[83] Iacobacci I, Pilehvar M T, Navigli R. SensEmbed: Learning sense embeddings for word and relational similarity//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 95-105

[84] Ni Yuan, Xu Qiong Kai, Cao Feng, et al. Semantic documents relatedness using concept graph representation//Proceedings of the 9th ACM International Conference on Web Search and Data Mining. San Francisco, USA, 2016: 635-644

[85] Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms//Proceedings of the 10th International Conference on World Wide Web. Hong Kong, China, 2001: 285-295

[86] Linden G, Smith B, York J. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7(1): 76-80

[87] Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems, 2002, 20(4): 422-446

[88] Davidson J, Liebald B, Liu J, et al. The YouTube video recommendation system//Proceedings of the 2010 ACM Conference on Recommender Systems. Barcelona, Spain, 2010: 293-296

[89] Graepel T, Candela J Q, Borchert T, Herbrich R. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's bing search engine//Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel, 2010: 13-20

[90] Ponnuswami A K, Pattabiraman K, Wu Q, et al. On composition of a federated Web search result page: Using online users to provide pairwise preference for heterogeneous verticals//Proceedings of the 4th International Conference on Web Search and Web Data Mining. Hong Kong, China, 2011: 715-724

[91] Kohavi R, Deng A, Frasca B, et al. Trustworthy online controlled experiments: Five puzzling outcomes explained//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 786-794

[92] Li Y, Hsu B-J P, Zhai C X. Unsupervised identification of synonymous query intent templates for attribute intents//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco, USA, 2013: 2029-2038

[93] Xie Ruobing, Liu Zhiyuan, Jia Jia, et al. Representation learning of knowledge graphs with entity descriptions//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 2659-2665

[94] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, USA, 2010: 1306-1313

[95] Herlocker J L, Konstan J A, Riedl J. Explaining collaborative filtering recommendations//Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. Philadelphia, USA, 2000: 241-250

[96] Bilgic M, Mooney R J. Explaining recommendations: Satisfaction vs. promotion//Proceedings of the Beyond Personalization 2005: A Workshop at the International Conference on Intelligent User Interfaces. San Diego, USA, 2005: 1-8

[97] Tintarev N, Masthoff J. A survey of explanations in recommender systems//Proceedings of the 23rd International Conference on Data Engineering Workshops. Istanbul, Turkey, 2007: 801-810

[98] Cramer H, Evers V, Ramlal S, et al. The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction, 2008, 18(5): 455-496

[99] Tintarev N, Masthoff J. Designing and evaluating explanations for recommender systems. Recommender Systems Handbook. Boston, MA, USA: Springer, 2011: 479-510

[100] Gedikli F, Jannach D, Ge M. How should I explain? A comparison of different explanation types for recommender systems. International Journal of Human-Computer Studies, 2014, 72(4): 367-382

[101] Tintarev N, Masthoff J. Explaining recommendations: Design and evaluation. Recommender Systems Handbook. Boston, MA, USA: Springer, 2015: 353-382

[102] Zhang Yongfeng. Explainable recommendation: Theory and applications. Tsinghua University, Beijing, China, 2016

[103] Tintarev N, Masthoff J. Evaluating the effectiveness of explanations for recommender systems. User Modeling and User-Adapted Interaction, 2012, 22(4-5): 399-439

[104] Vig J, Sen S, Riedl J. Tagsplanations: Explaining recommendations using tags//Proceedings of the 14th International Conference on Intelligent User Interfaces. Sanibel Island, USA, 2009: 47-56

[105] Ling X, Weld D S. Fine-grained entity recognition//Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Canada, 2012: 94-100

[106] Yosef M A, Bauer S, Hoffart J, et al. Hyena: Hierarchical type classification for entity names//Proceedings of the 24th International Conference on Computational Linguistics. Mumbai, India, 2012: 1361-1370

[107] Ren Xiang, He Wenqi, Qu Meng, et al. AFET: Automatic fine-grained entity typing by hierarchical partial-label embedding//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 1369-1378

[108] Abhishek A, Anand A, Awekar A. Fine-grained entity type classification by jointly learning representations and label embeddings//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017: 797-807

[109] Sharma A, Cosley D. Do social explanations work?: Studying and modeling the effects of social explanations in recommender systems//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 1133-1144

[110] Passant A. Dbrec-music recommendations using DBpedia//Proceedings of the 9th International Semantic Web Conference. Shanghai, China, 2010: 209-224

[111] Catherine R, Mazaitis K, Eskenazi M, Cohen W. Explainable entity-based recommendations with knowledge graphs//Proceedings of the Poster Track of the 11th ACM Conference on Recommender Systems. Como, Italy, 2017: 1-2

[112] Zhang Yongfeng, Lai Guokun, Zhang Min, et al. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast, Australia, 2014: 83-92

[113] Chen Li, Wang Feng. Explaining recommendations based on feature sentiments in product reviews//Proceedings of the 22nd International Conference on Intelligent User Interfaces. Limassol, Cyprus, 2017: 17-28

[114] Fang L, Sarma A D, Yu C, Bohannon P. REX: Explaining relationships between entity pairs. Proceedings of the VLDB Endowment, 2011, 5(3): 241-252

[115] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks//Proceedings of the Neural Information Processing Systems Conference. Montreal, Canada, 2014: 3104-3112

[116] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015

[117] Luong M-T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1412-1421

[118] Vinyals O, Kaiser L, Koo T, et al. Grammar as a foreign language//Proceedings of the Neural Information Processing Systems Conference. Montreal, Canada, 2015: 2773-2781

[119] Luong M-T, Sutskever I, Le Q V, et al. Addressing the rare word problem in neural machine translation//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China, 2015: 11-19

[120] Vinyals O, Fortunato M, Jaitly N. Pointer networks//Proceedings of the Neural Information Processing Systems Conference. Montreal, Canada, 2015: 2692-2700

[121] Gulcehre C, Ahn S, Nallapati R, et al. Pointing the unknown words//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 140-149

[122] Gu Jiatao, Lu Zhengdong, Li Hang, et al. Incorporating copying mechanism in sequence-to-sequence learning//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1631-1640

[123] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, 2017: 1073-1083

[124] Mi H, Sankaran B, Wang Z, Ittycheriah A. Coverage embedding models for neural machine translation//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 955-960

[125] Tu Zhaopeng, Lu Zhengdong, Liu Yang, et al. Modeling coverage for neural machine translation//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 76-85

[126] Turner J, Charniak E. Supervised and unsupervised learning for sentence compression//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. University of Michigan, USA, 2005: 290-297

[127] McDonald R. Discriminative sentence compression with soft syntactic evidence//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy, 2006: 297-304

[128] Galley M, McKeown K. Lexicalized Markov grammars for sentence compression//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Rochester, USA, 2007: 180-187

[129] Nomoto T. Discriminative sentence compression with conditional random fields. Information Processing & Management, 2007, 43(6): 1571-1587

[130] Galanis D, Androutsopoulos I. An extractive supervised two-stage method for sentence compression//Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, USA, 2010: 885-893

[131] Che Wanxiang, Zhao Yanyan, Guo Honglei, et al. Sentence compression for aspect-based sentiment analysis. Audio, Speech, and Language Processing, 2015, 23(12): 2111-2124

[132] Filippova K, Alfonseca E, Colmenares C A, et al. Sentence compression by deletion with LSTMs//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 360-368

[133] Cohn T, Lapata M. An abstractive approach to sentence compression. ACM Transactions on Intelligent Systems and Technology, 2013, 4(3): 41:1-41:35

[134] Yu Naitong, Zhang Jie, Huang Minlie, Zhu Xiaoyan. An operation network for abstractive sentence compression//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, USA, 2018: 1065-1076

[135] Mallinson J, Sennrich R, Lapata M. Sentence compression for arbitrary languages via multilingual pivoting//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 2453-2464

[136] Klerke S, Goldberg Y, Sogaard A. Improving sentence compression by learning to predict gaze//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA, 2016: 1528-1533

[137] Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: A method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318

[138] Lin Chin-Yew. ROUGE: A package for automatic evaluation of summaries//Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. Barcelona, Spain, 2004: 74-81

[139] Callison-Burch C, Fordyce C, Koehn P, et al. (Meta-) evaluation of machine translation//Proceedings of the 2nd Workshop on Statistical Machine Translation. Prague, Czech Republic, 2007: 136-158

[140] Wu Bin, Xiong Chenyan, Sun Maosong, Liu Zhiyuan. Query suggestion with feedback memory network//Proceedings of the 2018 World Wide Web Conference on World Wide Web. Lyon, France, 2018: 1563-1571

[141] Ahmad W, Chang Kai-Wei, Wang Hongning. Multi-task learning for document ranking and query suggestion// Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-14

[142] Amat F, Chandrashekar A, Jebara T, Basilico J. Artwork personalization at Netflix//Proceedings of the 12th ACM Conference on Recommender Systems. Vancouver, Canada, 2018: 487-488

[143] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data//Proceedings of the Neural Information Processing Systems Conference. Lake Tahoe, USA, 2013: 2787-2795

[144] Wang Zhen, Zhang Jianwen, Feng Jianlin, Chen Zheng. Knowledge graph embedding by translating on hyperplanes //Proceedings of the 28th AAAI Conference on Artificial Intelligence. Quebec City, Canada, 2014: 1112-1119

[145] Lin Yankai, Liu Zhiyuan, Sun Maosong, et al. Learning entity and relation embeddings for knowledge graph completion //Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2181-2187

[146] Ji Guoliang, Liu Kang, He Shizhu, Zhao Jun. Knowledge graph completion with adaptive sparse transfer matrix// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, Arizona, USA, 2016: 985-991

HUANG Ji-Zhou, Ph. D. candidate. His research interests include natural language processing, recommender system, and artificial intelligence.

SUN Ya-Ming, Ph. D., engineer. Her research interests include entity disambiguation and natural language processing.

WANG Hai-Feng, Ph. D., professor. His research interests include natural language processing, machine translation, and artificial intelligence.

LIU Ting, Ph. D., professor. His research interests include artificial intelligence, natural language processing, and social computing.

Background

Entity recommendation aims to provide search users with entity suggestions relevant to their information needs, which can help them to explore and discover entities of interest. For this reason, over the past few years, major commercial Web search engines have proactively recommended related entities for a query along with the regular Web search results to enrich and improve the user experience of information retrieval and discovery. The task of building an entity recommendation system presents more challenges than the task of building a traditional item-based recommender system because of the ambiguity of the entities mentioned in queries, the domain-agnostic recommendation methods for Web-scale queries, and the cross-domain recommendation scenarios. To address these challenges, the following three sub-tasks should be studied on building an entity recommendation system in Web search engines. The first is entity linking in queries, which aims to disambiguate the entity mentioned in a query and link it to the corresponding entity in a knowledge base. The second is entity recommendation, which aims to find a set of related entities to a query, and then rank these entities. The third is recommendation captioning, which aims to explain why two entities are related and why a group of entities is recommended to a user. However, the above problems have not been well addressed and remain considerable challenges.

Existing entity recommendation methods can be categorized into two groups: context-insensitive methods and context-aware methods. A major limitation of the existing context-insensitive methods is that the recommendations are generated based solely on the query, without considering any context information. As a result, they may fail to generate satisfactory entity recommendations because the query itself is insufficient to understand the information needs of users. By contrast, context-aware methods take into account the query as well as the context in generating recommendations. Recently, several studies have tried to provide users with context-aware entity recommendations. It has been shown that the performance can be significantly improved due to the help of context information. The existing methods can be further categorized into two types: personalized methods and global methods. The only difference between them is that personalized methods generate entity recommendations based on the query as well as a user's interests and preferences, while global methods ignore the user dimension.

This paper summarizes the previous related studies on entity recommendation system in Web search engines. The research background and the challenges of this task are presented first, and then the related studies are introduced. Finally, problems are discussed, and several future research directions are suggested.