Computer Science and Application 计算机科学与应用, 2019, 9(8), 1489-1499
Published Online August 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.98167

Hans汉斯

Fine-Grained Literature Retrieval Based on Structuring Keywords

Wei Yu[1], Peng Dai[2], Jie Zhang[2], Wenjun Qi[1]

[1]Jiangsu Frontier Electric Technology Co., Ltd., Nanjing Jiangsu
[2]Department of Computer Science and Engineering, Southeast University, Nanjing Jiangsu
Email: daipeng@seu.edu.cn, 15905166664@163.com

Received: Jul. 25th, 2019; accepted: Aug. 5th, 2019; published: Aug. 12th, 2019

Abstract

Intelligent retrieval requires intent recognition of the user's query. In the field of scientific literature retrieval, the user's potential query intent can be divided into problem-oriented and method-oriented. In this paper, we use the problem-oriented and method-oriented as the intent template, and propose a method to structure the query keywords by using the entity information, then to analyze and match the query intent. Specifically, named entity recognition technology is used to extract the entity and entity type information, express the query intent, and use the Markov Random Field graph model to model the query, query the joint probability of the entity and the document, and perform matching. The experimental results show that the structuring of keywords can effectively model the user's query intent thus giving more accurate, fine-grained search results for different query intents.

Keywords

Document Retrieval, Knowledge Base, Dependency Modeling, Fine-Grained

结构化关键词的细粒度文献检索

喻　伟[1]，戴　鹏[2]，张　杰[2]，戚文君[1]

[1]江苏方天电力技术有限公司，江苏　南京
[2]东南大学计算机科学与工程学院，江苏　南京
Email: daipeng@seu.edu.cn, 15905166664@163.com

收稿日期：2019年7月25日；录用日期：2019年8月5日；发布日期：2019年8月12日

摘　要

智能化的检索需要对用户的查询进行意图识别，在科学文献检索领域，用户的潜在查询意图可分为面向

文章引用: 喻伟, 戴鹏, 张杰, 戚文君. 结构化关键词的细粒度文献检索[J]. 计算机科学与应用, 2019, 9(8): 1489-1499.
DOI: 10.12677/csa.2019.98167

问题和面向方法。本文以问题和方法为意图模版，提出一种利用实体信息对查询关键词进行结构化，进行查询意图解析及匹配的方法。具体为使用命名实体识别技术抽取实体及实体类型信息，表达查询意图，并利用马尔可夫随机场图模型建模查询、查询实体与文献的联合概率，进行匹配。实验结果表明，对关键词进行结构化能有效从上述两个角度建模用户的查询意图，从而对于不同的查询意图能够给出更精确、细粒度的检索结果。

关键词

文献检索，知识图谱，依赖建模，细粒度

1. 引言

随着文献数量的急剧增长，信息检索技术已经成为广大科研工作者和技术从业者获取知识的一个重要途径。现代搜索引擎技术有两个核心概念：智能程度和自然程度。智能程度是指理解用户的意图和文档内容，然后快速、准确的找出相关答案。自然程度是指根据用户输入的搜索请求，把搜索结果自然地、清晰的呈现给用户。提升搜索引擎的智能程度和自然程度，具有提升用户检索效率、使用体验，增加系统粘性的作用。

传统的文献搜索引擎通过对文献数据进行全文索引，基于关键字匹配在语料集中进行检索。由于只涉及词语的匹配，所以检索效率高，但是缺点也很明显，完全依赖用户的查询词，缺少语义分析和扩展能力，难以保证较好的检索性能。文献检索还具有多实体和查询面向问题－方法两个特点。如查询“deep learning for information retrieval”，就包含“deep learning”和“information retrieval”两个概念实体。传统的基于关键字匹配的方法难以准确理解科学概念实体含义，这是造成查询误差的一个重要因素。因此，基于知识库的语义检索近年来成了文献检索的一个重要研究方向。然而，这些工作在利用知识库时没有考虑到查询实体间的内在联系，仅将实体作为单字补充信号改善检索效果。这样会导致检索结果更容易倾向于包含单个高频实体的文档，而不是同时包含多个不同实体的文档。另一方面，科学文献检索中用户信息需求可以分为两类：1) 面向问题，即查询某个科学领域问题、该问题的引申；2) 面向方法，即不限定具体的科学问题，而是查询某一类技术方法的应用。现有框架均没有考虑到两种检索需求的区别，需要用户在检索结果中进行二次查询或人工甄别。

为了解决上述问题，本文提出了一个基于实体依赖建模的检索模型。模型在[1]基础上改进，对用户查询语句中的实体间的依赖关系进行了建模。同时，为了对用户的查询意图进行解析，对于文档，我们抽取了类型信息。实体类型反映了实体在文档语境下表述的是问题还是方法信息。由于查询中实体类型难以得到，本文提出一种基于伪相关反馈方法对用户查询类型进行估计。通过解析出用户的问题方法和需求表述，能够从上述两个角度解释用户查询意图，从而对于不同的查询意图能够给出更加精确、细粒度的检索结果。

2. 相关工作

2.1. 学术搜索引擎

文献检索的现实需求推动了许多学术搜索引擎的发展。CiteSeerX [2]，ArnetMiner [3]，PubMed [4]，

Microsoft Academic Search (MAS) [5]等文献搜索引擎中大量应用数据挖掘技术、自动信息抽取技术、推荐技术改善用户检索体验。如 MAS 通过建立文献知识图谱，可以在检索结果页给出查询相关研究主题及领域相关作者，提供更丰富的结果展示。然而上述大部分系统的研究侧重于学术数据的分析任务，例如文献知识图构建[5]，文献重要性建模[6]，文档摘要[7]，研究者社群关系[3]等等，科学文献的检索算法仍具有很大的研究价值。

2.2. 基于实体的检索模型

实体(例如人，位置或抽象概念)是用于组织和检索信息的最小自然单元。有研究发现，超过 70%的 Bing 查询和超过 50%的 Semantic Scholar 文献查询与实体相关[8] [9]。随着知识图谱相关技术的逐渐成熟，越来越多的工作研究如何利用实体信息改进检索算法。

一种直觉的方法是利用外部知识库中实体相关信息(实体描述，属性等)去增强用户查询，也即查询扩展。He 等[10]将维基百科中的实体描述用作伪相关反馈语料库以获得更清晰的扩展术语。Dalton 等[11]使用查询相关实体属性的文本字段扩展查询，并基于扩展文本生成更丰富的排序特征。另一类利用实体的方法是试图将查询和文档都映射到同一个实体空间进行相似度比较。Liu 和 Fang [12]使用查询及文档中的实体来构造潜在的实体空间，然后在高维实体空间内计算查询映射和文档映射的相关度。最近的趋势是构建基于实体的文本表示，并将其与传统的基于单词的表示相结合进行检索。Xiong 等[9]提出使用实体向量来计算查询和文档的相似度，以提高基于单词的检索模型效果。

3. 基于实体依赖建模的检索模型

本节将详细介绍我们提出的基于实体依赖的检索模型。如前文所说，本模型是[1]提出的 MRF 检索模型的一个扩展，因此我们先简要介绍 MRF 检索模型，再介绍我们的扩展方法。

3.1. MRF 检索模型

马尔可夫随机场是无向图模型，它提供了一种紧凑，稳健的联合分布建模方法。在信息检索领域，常常希望建模查询 $Q=q_1,q_2,\cdots,q_n$ 与文档 D 的联合概率。MRF 检索模型假设查询和相关文档组成的<查询，文档>对存在一个潜在分布，也即从这个分布中采样出来的查询和文档都是相关的。MRF 检索模型则是建模查询 Q 和文档 D 的相关性。

一个马尔可夫随机场是由一个无向图 G，以及定义在图 G 的团(clique)上一系列势函数(potential function)组成。其中图上的节点表示随机变量，边表示变量间的依赖性。一个马尔可夫随机场需要满足马尔可夫性质，即网络中每个节点 v 都条件独立于其邻居节点给定时的 v 的任意非邻居节点子集。

给定一个图 G，以及势函数 ψ 和参数向量 Λ，则定义在 Q 和 D 上联合概率为：

$$P_{G,\Lambda}(Q,D)=\frac{1}{Z_\Lambda}\prod_{c\in C(G)}\psi(c;\Lambda)$$

其中 Z_Λ 是归一化因子，为了便于计算势函数 ψ 通常取指数函数 $\exp\{\lambda_i f_i(c)\}$，$f_i(c)$ 则是定义在团 c 上的是实值特征函数。

对于给定查询 Q 如何与文档 D 构建无向概率图，原始检索模型给出了三种方式：完全独立(fully term independency)，序列依赖(sequential term dependency)和完全依赖(full term dependency)。分别对应于查询项之间完全独立，序列依赖，以及完全依赖的情况，如图 1：

MRF 模型的联合概率分布通常基于无向图上的最大团参数化，但用于检索模型时，这样的方式太过粗糙。这里为了在更细粒度的水平上将特征函数与团相关联，同时保持特征的数量，从而保持参数的数

量。对于上述定义的三种类型的团，每个定义一种特征函数。

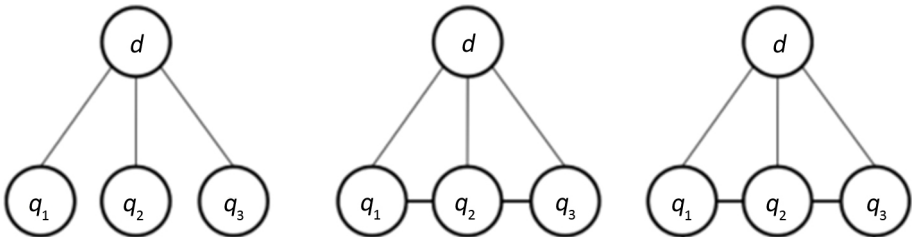

Figure 1. Three dependency modeling approaches
图 1. 三种依赖建模方式

3.2. 检索框架

上一节节中我们简要介绍了 MRF 模型，本模型是在 MRF 模型的基础上，引入实体信息的一种扩展模型。

原始的 MRF 模型主要是对查询词的三种不同类型的依赖关系进行建模，即完全独立、序列依赖和完全依赖，分别对应于计算图 $\langle Q,D\rangle$ 上三种不同类型团的特征函数。在本模型中，我们加入一种新的节点类型-查询实体节点 $E=\{e_1,e_2,\cdots\}$，并对由 $\{Q,D,E\}$ 构成的概率图 G 计算联合概率。如图 2 为加入实体集 E 后的 MRF 结构，查询实体 E 是由多个实体 $(e_1,e_2,e_3,\cdots)$ 组成，由查询经过实体链接技术抽取得到，在这里查询实体和查询词两部分的内部元素间采用相互独立关系。

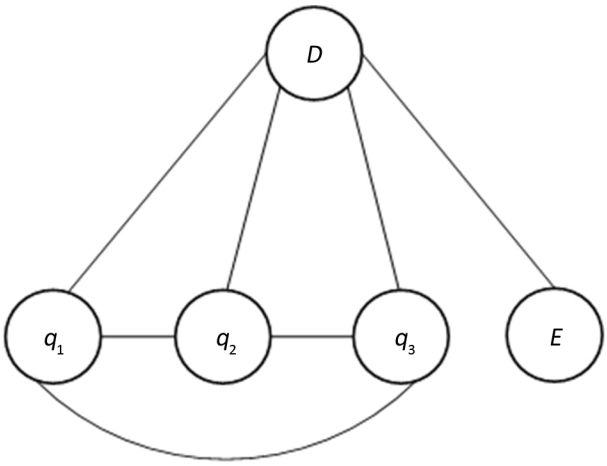

Figure 2. MRF probabilistic undirected graph containing query entities
图 2. 包含查询实体的 MRF 概率无向图

依据 MRF 原理，对于包含 $\langle D,Q,E\rangle$ 的概率无向图 G，我们类似的定义查询 Q、文档 D 和查询实体 E 的联合概率为：

$$P_{G,\Lambda}(D,Q,E)=\frac{1}{Z_\Lambda}\prod_{c\in C(G)}\psi(c,\Lambda)$$

最终我们使用条件概率 $P_{G,\Lambda}(D,Q,E)$ 作为最后的文档检索得分：

$$P_{G,\Lambda}(D\mid Q,E)=\frac{P_{G,\Lambda}(Q,D,E)}{P_{G,\Lambda}(Q,E)}\Rightarrow \log P_{G,\Lambda}(Q,D,E)-\log P_{G,\Lambda}(Q,E)\Rightarrow \sum_{c\in C(G)}\log\psi(c;\Lambda) \quad (1)$$

在这里我们可以把团分为两个大的部分。第一部分为文档 D 与查询词 Q 之间的团集 $C_w(G)$，即满足前文所述的查询词完全独立、序列依赖、完全依赖假设的团集。第二部分为文档 D 与查询实体之间的团集 $C_E(G)$，这一部分的团集类型以及对应的特征函数将在 3.3 和 3.4 节详细说明。相应的，我们将公式简化为如下：

$$P_{G,\Lambda}(D|Q,E)=(1-\lambda_E)\sum_{c\in C_w(G)}\log\psi_w(c;\Lambda)+\lambda_E\sum_{c\in C_E(G)}\log\psi_E(c;\Lambda) \quad (2)$$

3.3. 实体依赖模型

参考查询词依赖建模的方式，对于查询实体的依赖建模也可以遵循上述不同的依赖假设，即完全独立、序列依赖、完全依赖假设。对应于上述假设，分别定义三种团集，如表 1：

Table 1. Entity dependency type definition
表 1. 实体依赖类型定义

团集类型	说明
T_E	包含文档节点和一个查询实体节点的团集。
O_E	包含文档节点和在查询中出现的多个(两个及以上)连续查询实体的团集。
U_E	包含文档节点和在查询中以任意顺序出现的多个(两个及以上)查询实体的团集。

对查询实体建模需要考虑到实体的特性，例如每个查询实体本身就是一个完整的语义单元，每个实体可以是单字、双字或者多字短语等。如图 3，查询经过实体链接工具标注后得到 4 个不同实体 e_1,e_2,e_3,e_4，虽然在原始查询中它们并不是连续的，但我们认为在实体层面上它们是连续的序列。同样的可以得到文档实体序列。那么依据表 1 中定义，有 $T_E=\{(e_1,en_1),(e_2,en_2),(e_4,en_5)\}$，$O_E=\{(e_1,e_2,en_1,en_2)\}$，$U_E=\{(e_1,e_2,en_1,en_2),(e_1,e_4,en_1,en_5),(e_2,e_4,en_2,en_5)\}$ 为只考虑两个实体时的完全依赖匹配团集。在本模型中我们只使用完全独立和完全依赖假设，而忽略了对于序列依赖关系的建模。下面将结合实例对上述选择进行详细阐述。

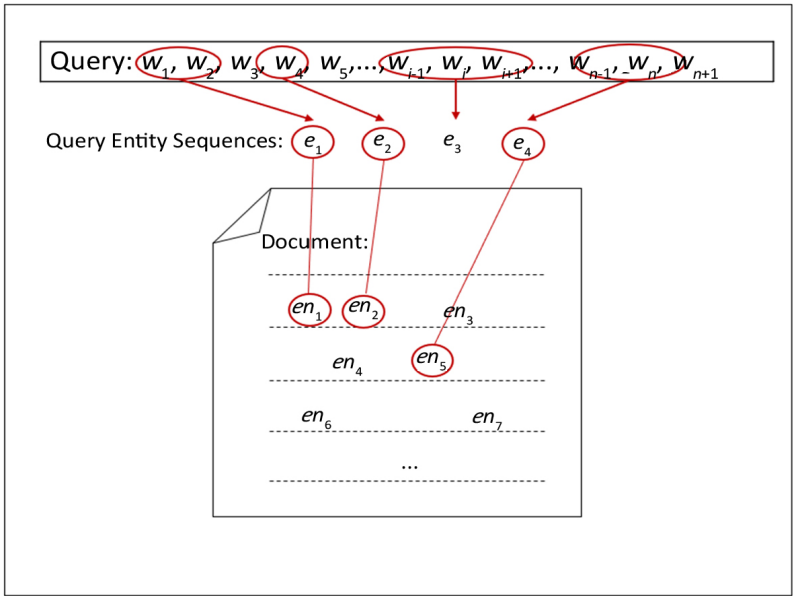

Figure 3. Examples of query and document entity clique set matching
图 3. 查询和文档实体 clique 集匹配示例

在科学文献检索领域，除却题名、作者和研究机构等相关字段的简单查找，用户需要根据自己的检索需求提取关键字，然后组合形成最终查询。如查询“基于深度学习的信息检索”。其中“深度学习”和“信息检索”两个概念实体，表述了用户不同的信息需求。直观上，这两个实体在文档中出现次数越高则文档越相关。T_E 集合上计算的正是文档对不同信息需求匹配的程度。然而单纯依赖于 T_E 集合上的结果进行排序，会导致检索结果更容易倾向于包含单个高频实体的文档，而不是同时包含多个不同实体的文档。

O_E 和 U_E 都体现了实体的共现关系，但这里不使用 O_E 团集，这是基于以下两个因素的考虑：1) 由于实体本身为一个独立语义单元，实体与实体之间序列依赖关系并不明显，一个常见的例子是对于多关键词罗列的查询，如“CRISPR/Cas9，ZFN，TALEN”，替换关键词之间的顺序，并不影响查询结果。2) 受限于实体识别精度，查询中识别出的实体并不完备，实体间可能有未识别出的实体。

3.4. 特征函数定义

目前为止，我们在由 $\{D,Q,E\}$ 组成的无向图上定义了 5 种不同的团集，其中 T、O 和 U 与原始 MRF 模型一样，特征函数也相同，这里不再赘述，O_E 和 U_E 为新增团集。

文献检索场景下，返回的文档(即文献)包含了标题、摘要、关键词、正文等多个字段。我们假设一个文档 d_i 有 k 个不同字段，$d_i=\{d_i^1,d_i^2,\cdots,d_i^k\}$，那么文档集合就可以被分成 k 部分：$\{D_1,D_2,\cdots,D_k\}$。我们使用混合语言模型(mixture of language model, MLM) [13]定义特征函数。

对于 T_E 团集上的特征函数 $f_{T_E}(D,E)$，我们定义：

$$f_{T_E}(e_i,d_j)=\log P(e_i\mid d_j)=\log\sum_{m=1}^{k}w_m^T\cdot P(e_i\mid d_j^m)$$

其中 $m=\overline{1,k}$，k 是文档包含的字段的总数目，w_m^T 是每一个字段对应的权重，概率 $P(e_i\mid d_j^m)$ 表示每个文档字段下实体生成概率，通过极大似然估计得到：

$$P(e_i\mid d_j^m)=\frac{n_{e_i,d_j^m}+u_m^T\dfrac{n_{e_i,D_m}}{L_{D_m}}}{L_{d_j^m}+u_m^T}$$

n_{e_i,d_j^m} 表示实体 e_i 出现在 d_j^m 的次数，$L_{d_j^m}$ 表示文档 j 字段 m 的总长度。与 n_{e_i,d_j^m} 和 $L_{d_j^m}$ 定义类似，n_{e_i,D_m} 和 L_{D_m} 为在文档集合上的定义，u_m^T 是字段 j 的狄利克雷平滑因子[14]。

至于 U_E 集合上，出于对计算效率的考虑，本文只考虑两个实体的情况，定义特征函数如下：

$$f_{U_E}(e_{i,i+1},d_j)=\log P(e_i,e_{i+1}\mid d_j)=\log\sum_{m=1}^{k}w_m^U\cdot P(e_i,e_{i+1}\mid d_j^m)$$

对于 $P(e_i,e_{i+1}\mid d_j^m)$ 我们统计两个实体在同一个窗口单元(如同一字段)共现的频次，即：

$$P(e_i,e_{i+1}\mid d_j^m)=\frac{n_{e_{i,i+1},d_j^m}+u_m\dfrac{n_{e_{i,i+1},D_m}}{L_{D_m}}}{L_{d_j^m}+u_m}$$

在 MRF 基础模型中对于每个匹配项，假设具有相同的权重。这样的假设可能对检索性能有潜在的不利影响，特别是对于复杂的冗长查询[15]。对此，我们引入了一个权重项 $g(\cdot)$，用于度量不同项的重要程度。这里我们参考 IDF 定义方式，定义权重函数 $g(\cdot)$：

$$g(t)=1+\log\left(\frac{N_D}{n_t}\right)$$

其中 t 可以是 unigram 或者 bigram，N_D 表示文档集合数目，n_t 表示 t 出现的文档个数。最终特征函数为：

$$f_{T_E}(e_i,d_j)=g(e_i)\cdot\log\sum_{m=1}^{k}w_m^T\cdot P(e_i\mid d_j^m) \tag{3}$$

$$f_{U_E}(e_{i,i+1},d_j)=g(e_i,e_{i+1})\cdot\log\sum_{m=1}^{k}w_m^U\cdot P(e_i,e_{i+1}\mid d_j^m) \tag{4}$$

3.5. 实体类型信息

本文关注文献检索领域，依据文献检索的特点将用户的检索查询划分为面向问题和面向方法两种信息需求。为了实现上述检索方式，我们在检索模型中引入实体的类型信息。通过命名实体识别(Named Entity Recognition，简称 NER)技术，首先识别出文本中具有特定意义的实体。

本文中涉及的实体类别主要为以下几种类别：

- Task (T)：表示与具体应用，最终目标或问题定义相关的科学概念实体。
- Process (P)：表示与某些科学模型，算法或过程相关的科学概念实体。
- Other (O)：表示除上述定义的两种类型实体外，文献中涉及的其它科学概念实体。

本文使用模型 CNN-biLSTM-CRF [16]进行特定域命名实体识别。如图 4 为典型的一篇研究文献摘要的实体指称抽取过程。

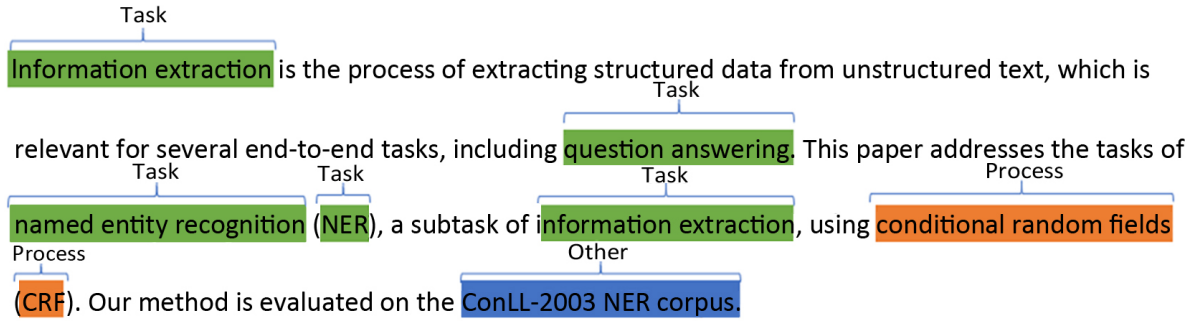

Figure 4. Examples of specific type entity extraction
图 4. 特定类型实体抽取示例

biLSTM-CRF 模型在进行 NER 任务时，利用上下文信息，编码了文本的词法和语法信息，因此处理短句子或者不完整文本描述效果并不好，适合于文献标题、摘要、正文等长文本数据的实体识别。对于查询上实体类别信息的识别，本文基于伪相关反馈策略对查询中实体类型进行估计。具体的，首先利用公式(2)计算与用户查询的相关文档，然后基于相关文档，我们定义查询实体类型的概率分布为：

$$P(t\mid E,D,Q)\propto P(t\mid D)P(D\mid Q,E)\approx\sum_{d_i\in D_R}P(t\mid d_i)P(d_i\mid Q,E)$$

D_R 表示与查询相关的 TOP-K 的相关文档集合，t 表示实体类型，$t\in\{T,P,O\}$。

然后，我们使用上式对于用户查询实体类型进行估计，得到用户查询类型推断，基于用户查询类型进行再次检索。具体的做法是，我们对公式(3) (4)稍加改变，引入实体的类别信息。本文中采用以下策略，在估计语言模型概率 $P(e_i\mid d_j^m)$ 和 $P(e_i,e_{i+1}\mid d_j^m)$ 时，我们考虑词项的类型匹配信息。对于词项 e (即实体、实体对)在 d_j^m 出现的词频 n_{e,d_j^m} 估计有：

$$n_{e_i,d_j^m} = \sum_{e_i \in E_q} \sum_{e_j \in E_d} match\left(e_i, e_j\right)$$

$$match\left(e_i, e_j\right) = \begin{cases} 1 & type\left(e_i\right) \neq type\left(e_j\right) \wedge e_i = e_j \\ 0 & e_i \neq e_j \\ \eta & type\left(e_i\right) = type\left(e_j\right) \wedge e_i = e_j \end{cases}$$

其中η为超参数，值越大表示实体的类型信息越重要。最后使用公式(2)进行检索，得到满足用户查询需求的文献结果。

如图 5，为基于实体依赖建模的 MRF 检索流程。其主要分为离线和在线部分，离线部分(左)负责文献集合的实体抽取和索引构建。在线部分(右)可以总结为以下流程：1) 用户根据自身潜在信息需求，形成原始查询；2) 原始查询经过实体链接，变成查询以及查询实体，然后将查询以及查询实体送入检索模型，使用公式(2)进行初次检索，得到 TOP-K 的相关文档；3) 基于相关文档中实体类型，估计查询实体的类型分布，获得用户的查询类型信息。4) 基于查询词、查询实体和实体类型信息再次进行面向问题-方法的细粒度检索。

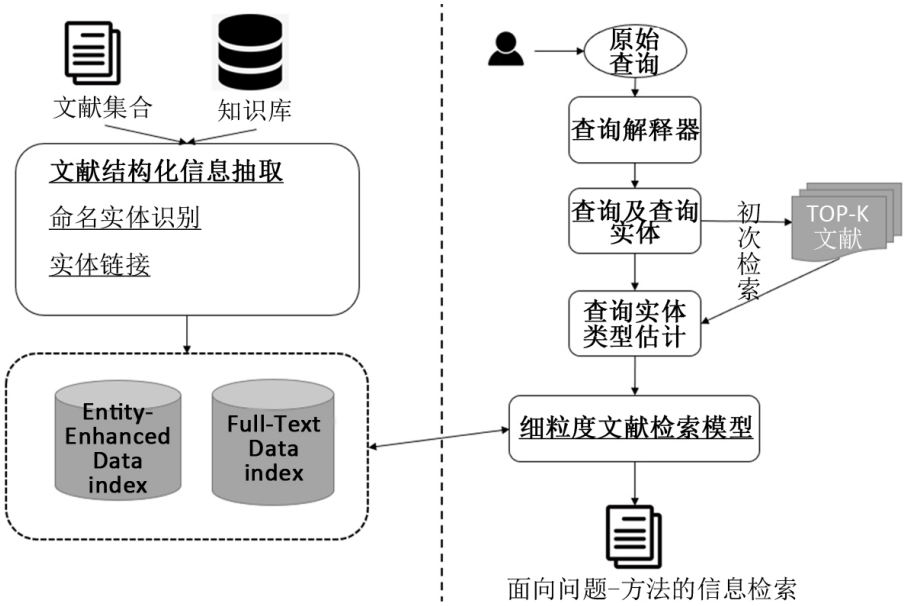

Figure 5. MRF retrieval process based on entity dependency modeling
图 5. 基于实体依赖建模的 MRF 检索流程

4. 实验

4.1. 实验数据集

本文使用 Explicit Semantic Ranking (ESR) [9]作为我们的基准数据集，验证我们提出的基于实体依赖建模的检索模型的有效性。

ESR 数据集的查询是从学术搜索引擎 Semantic Scholar 在 2016 的查询日志中采样构成。总共包含 100 个查询，其中 20 个采样自高频查询，30 来自中频查询，剩余 50 个采样自 Semantic Scholar 表现较差的困难查询，主要涉及计算机科学领域。查询候选文档是从 Semantic Scholar 在线搜索系统汇集生成，共有 8541 篇候选文档，每个相关文档都由人工标记，共分 5-level 的相关度。

4.2. 实体识别

实体抽取部分，分为查询和文献语料两个部分。对于文献语料，我们首先需要在 ESR 数据集上抽取特定类型(Task, Process)的实体指称。这里使用在 ScienceIE SemEval 2017 Task 10 [17]数据集上训练得到的 CNN-BiLSTM-CRF 命名实体识别模型。然后使用开源实体链接工具 dexter2 [18]将 Bi-LSTM-CRF 识别的结果链接到 Wikipedia 上。对于用户查询，我们直接使用 dexter2 进行实体识别和链接。

经过实体链接后，100 个查询其中由 92 个查询包含实体，其中 32 个为多实体查询，最大实体个数为 3。ESR 文献语料集总共包含 179,702 个实体。

4.3. 检索性能

评价指标：我们使用信息检索中常用的 NDCG 和 Percision 作为我们主要的评价指标。

基准模型：这里我们选取两个常用的信息检索模型作为基准模型对比。其中包括 BM25 模型，查询语言模型(Query Likehood, QL)。所有的模型都被应用到文献的多个字段(标题，关键字，摘要)。

对于模型参数，所有模型均采用 5-折交叉验证进行参数调优。其中一折数据作为测试集，其余作为验证集。我们使用网格搜索(grid-search)方式在验证集上筛选最优的参数，使得 NDCG@20 最大化。使用网格搜索时，狄利克雷平滑因子 u (所有字段共用)筛选范围是{100, 500, 1000, 1500, 2000, 2500, 3000}；λ_E 取值范围[0,1]间隔 0.1；摘要、关键字和标题的权重取值范围是{1, 5, 10, 15, 20, 15, 30}；参数 η 取值范围是{1, 2, 3, 5, 7}。

Table 2. Comparison of BM25, QL and EBMRF retrieval results
表 2. BM25、QL 以及 EBMRF 的检索结果对比

	NDCG@20	P@20
BM25-W	0.3554	0.2977
BM25-E	0.3232	0.2742
BM25-Dual	**0.3686**	**0.3075**
QL-W	0.3497	0.2916
QL-E	0.3185	0.2612
QL-Dual	**0.3598**	**0.2998**
EBMRF-W	0.3735	0.3106
EBMRF-E	0.3499	0.2901
EBMRF-Dual	**0.3924**	**0.3315**

为了探讨实体信息有效性，我们在 BM25 模型、查询语言模型(QL)以及本文提出模型(EBMRF)上进行实验。实验只使用到了实体本身，没有使用实体类型信息。实验分为三个组别：1) 仅使用词信息；2) 仅使用实体信息；3) 同时使用词和实体信息。实验结果见表 2。从表中可以看出，三种模型仅使用单一信息的检索结果要低于使用两种信息的情况，说明引入实体信息时能够有效的提升检索性能。此外，还可以看出本文方法的最优检索性显著优于两种基准模型，本文分析是因为 EBMRF 方法充分利用了词项、查询实体间的依赖关系，而另外两种模型并没有考虑。

为了研究实体间关系对于提高检索性能的帮助，本文设置如下对照实验组，第一组为使用了全部 5 种类型的团的 EBMRF 模型；第二组为去掉实体依赖特征的 EBMRF 模型，记为 EBMRF+。在由 32 个多实体查询构成的子集 ESR-ME 以及 ESR 数据集上，结果如表 3。可以发现 EBMRF 模型较 EBMRF+取得

更高的检索性能，表明建模实体间关系有利于提升检索表现，并且在多实体查询集时，更加显著，从侧面说明捕捉实体依赖性对于提高多实体类型查询检索性能有效。

Table 3. Comparison of retrieval results between EBMRF and EBMRF+ on multi-entity queries
表 3. EBMRF 和 EBMRF+在多实体查询上检索结果对比

	ESR		ESR-ME	
	NDCG@20	P@20	NDCG@20	P@20
EBMRF	**0.3924**	**0.3315**	**0.4468**	**0.3761**
EBMRF^{+}	0.3819	0.3207	0.4201	0.3568

另一方面为了研究本文提出的面向问题方法的检索。我们设置对比实验，其中一组为使用了实体类型信息，记为 EBMRF*，一组为没有使用类型信息，记为 EBMRF。虽然 ESR 数据集在构建时并没有考虑到用户查询意图是基于问题还是方法，对相关文档的标注可能与本文研究的不完全相符，但我们仍好奇引入了类型信息之后是否会得到性能提升。实验结果如表所示。由表 4 可知，EBMRF*在 NDCG 和 P@{5, 10}上的结果要优于没有引入类型信息的 EBMRF，而在 NDCG@20 和 P@20 上提升不大。我们分析，产生这个结果的原因是类型信息的引入能够捕捉到更丰富的语义信息，使得高相关文档排名更靠前；而对于弱相关文档，其类型信息作用并不显著，而且因为多引入了实体识别的误差，导致性能持平或者下降。

Table 4. Comparison of retrieval results between EBMRF and EBMRF*
表 4. EBMRF 和 EBMRF*的检索结果对比

	ESR		
	NDCG@5	NDCG@10	NDCG@20
EBMRF	0.3245	0.3502	0.3924
EBMRF*	**0.3381**	**0.3586**	**0.3966**
	P@5	P@10	P@20
EBMRF	0.4483	0.3978	**0.3315**
EBMRF*	**0.4704**	**0.4046**	0.3288

5. 总结

在本文中，我们研究了科学文献领域的信息检索方法。根据文献查询特点，我们提出了一种新的面向问题－方法的检索方式。与传统检索不同的是，我们将用户的信息检索需求分为问题查询需求和方法查询需求，不同的文章侧重表达了用户的不同需求类型。为了捕捉用户的查询需求类型，我们提出了一种基于伪相关反馈的方法，利用相关文档中的实体类型，从而推测用户查询中的需求类型。另一方面，为了探索实体信息特别是实体间的内在关系的有效性，本文在 MRF 检索模型的基础上进行扩展，融合了实体和实体类型特征。实验表明引入实体补充信息以及建模实体间内在依赖关系对检索性能有提升。同时我们也发现引入用户信息需求类型对于检索也具有积极作用。

文献检索是近年来的热门研究方向，本文对这一领域的研究有一定的实用价值，但还有很多可以继续深入的地方。本文仅使用了二元实体对关系，实体间的关系可能更多元化；其次本文根据实体类型信息实现了面向问题方法的检索，类似地利用实体类型信息的检索方法值得进一步探索。

参考文献

[1] Metzler, D. and Croft, W.B. (2005) A Markov Random Field Model for Term Dependencies. *Proceedings of the* 28*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, 15-19 August 2005, 472-479. https://doi.org/10.1145/1076034.1076115

[2] Wu, J., William, K., Chen, H.H., *et al*. (2015) CiteSeerX: AI in a Digital Library Search Engine. *AI Magazine*, **36**, 35-48. https://doi.org/10.1609/aimag.v36i3.2601

[3] Tang, J., Zhang, J., Yao, L., *et al*. (2008) ArnetMiner: Extraction and Mining of Academic Social Networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, 24-27 August 2008, 990-998. https://doi.org/10.1145/1401890.1402008

[4] Lu, Z. (2011) PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. *The Journal of Biological Databases and Curation*, **2011**, baq036. https://doi.org/10.1093/database/baq036

[5] Sinha, A., Shen, Z., *et al*. (2015) An Overview of Microsoft Academic Service (MAS) and Applications. *Proceedings of the* 24*th International Conference on World Wide Web*, Florence, 18-22 May 2015, 243-246. https://doi.org/10.1145/2740908.2742839

[6] Shen, J., Song, Z., *et al*. (2016) Modeling Topic-Level Academic Influence in Scientific Literatures. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence Scholarly Big Data*: *AI Perspectives*, *Challenges*, *and Ideas*, Phoenix, AZ, 711-717.

[7] Ren, X., Shen, J., Qu, M., *et al*. (2017) Life-iNet: A Structured Network-Based Knowledge Exploration and Analytics System for Life Sciences. *Proceedings of ACL* 2017, *System Demonstrations*, Vancouver, July 2017, 55-60. https://doi.org/10.18653/v1/P17-4010

[8] Guo, J., Xu, G., Cheng, X., *et al*. (2009) Named Entity Recognition in Query. In: *Proceedings of the* 32*nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 267-274. https://doi.org/10.1145/1571941.1571989

[9] Xiong, C., Power, R. and Callan, J. (2017) Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding. *The* 26*th International World Wide Web Conferences Steering*, Perth, 3-7 April 2017, 1271-1279. https://doi.org/10.1145/3038912.3052558

[10] He, T. and Dai, X. (2013) Pseudo-Relevance Feedback Query Based on Wikipedia. *IEEE International Conference on Granular Computing*, Hangzhou, 11-13 August 2012, 1-6. https://doi.org/10.1109/GrC.2012.6468659

[11] Dalton, J., Dietz, L. and Allan, J. (2014) Entity Query Feature Expansion Using Knowledge Base Links. In: *Proceedings of the* 37*th International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM, New York, 365-374. https://doi.org/10.1145/2600428.2609628

[12] Liu, X. and Fang, H. (2015) Latent Entity Space: A Novel Retrieval Approach for Entity-Bearing Queries. *Information Retrieval Journal*, **18**, 473-503. https://doi.org/10.1007/s10791-015-9267-x

[13] Ogilvie, P. and Callan, J. (2003) Combining Document Representations for Known-Item Search. In: *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 143-150. https://doi.org/10.1145/860435.860463

[14] Zhai, C. and Lafferty, J. (2004) A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems* (*TOIS*), **22**, 179-214. https://doi.org/10.1145/984321.984322

[15] Bendersky, M., Metzler, D. and Croft, W.B. (2011) Parameterized Concept Weighting in Verbose Queries. In: *Proceedings of the* 34*th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 605-614. https://doi.org/10.1145/2009916.2009998

[16] Ma, X. and Hovy, E. (2016) End-to-End Sequence Labeling via Bi-Directional LSTM-CNNS-CRF. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, Berlin, 7-12 August 2016, 1064-1074. https://doi.org/10.18653/v1/P16-1101

[17] Augenstein, I., Das, M., Riedel, S., *et al*. (2017) SemEval 2017 Task 10: ScienceIE—Extracting Keyphrases and Relations from Scientific Publications. *Proceedings of the* 11*th International Workshop on Semantic Evaluations*, Vancouver, 3-4 August 2017, 546-555. https://doi.org/10.18653/v1/S17-2091

[18] Ceccarelli, D., Lucchese, C., Orlando, S., *et al*. (2013) Dexter: An Open Source Framework for Entity Linking. *International Workshop on Exploiting Semantic Annotations in Information Retrieval*, San Francisco, 28 October 2013, 17-20. https://doi.org/10.1145/2513204.2513212

知网检索的两种方式：

1. 打开知网首页：http://cnki.net/，点击页面中“外文资源总库 CNKI SCHOLAR”，跳转至：http://scholar.cnki.net/new，搜索框内直接输入文章标题，即可查询；
   或点击“高级检索”，下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询。
2. 通过知网首页 http://cnki.net/顶部“旧版入口”进入知网旧版：http://www.cnki.net/old/，左侧选择“国际文献总库”进入，搜索框直接输入文章标题，即可查询。

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org