软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(6):1584−1605 [doi: 10.13328/j.cnki.jos.005256]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

实体搜索综述*

张香玲[1,2], 陈跃国[1,2], 马登豪[1,2], 陈 峻[1,2], 杜小勇[1,2]

[1](数据工程与知识工程教育部重点实验室(中国人民大学),北京 100872)

[2](中国人民大学 信息学院,北京 100872)

通讯作者: 陈跃国, E-mail: chenyueguo@ruc.edu.cn

摘 要: 与传统的以网页页面集合的方式呈现搜索结果不同,实体搜索的结果是实体或实体集合,其优点是无需用户在纷杂的网页里面进行二次查找,更能提升用户的搜索体验.实体搜索的任务可以分为相关实体搜索和相似实体搜索.对近年来这两类任务的实体搜索技术进行综述.首先给出了实体搜索的形式化定义,并介绍了常用的评测指标;然后,对两种不同形式的实体搜索任务在两类数据源(非结构化数据集和结构化数据集)上的主要研究方法进行了详细的阐述和对比;最后,对未来的研究内容和发展方向进行了探讨和展望.

关键词: 实体搜索;对象搜索;相关实体搜索;相似实体搜索;知识图谱

中图法分类号: TP311

中文引用格式: 张香玲,陈跃国,马登豪,陈峻,杜小勇.实体搜索综述.软件学报,2017,28(6):1584−1605. http://www.jos.org.cn/1000-9825/5256.htm

英文引用格式: Zhang XL, Chen YG, Ma DH, Chen J, Du XY. Survey on entity search. Ruan Jian Xue Bao/Journal of Software, 2017,28(6):1584−1605 (in Chinese). http://www.jos.org.cn/1000-9825/5256.htm

Survey on Entity Search

ZHANG Xiang-Ling[1,2], CHEN Yue-Guo[1,2], MA Deng-Hao[1,2], CHEN Jun[1,2], DU Xiao-Yong[1,2]

[1](Key Laboratory of Data Engineering and Knowledge Engineering, MOE (Renmin University of China), Beijing 100872, China)

[2](School of Information, Renmin University of China, Beijing 100872, China)

Abstract: Entity search differs from traditional search engines in that the results of traditional search engines are Web pages, whereas the results of entity search are entities which can enhance the user's search experience. Entity search can be further categorized into the task of related entity search and the task of similar entity search. In this paper, a survey is presented on the techniques of entity search. Firstly, entity search is defined formally, and frequently used evaluation measures are introduced as well. Secondly, the algorithms of the two different types of entity search on two different data sources (unstructured data and structured data) are reviewed in details. Finally, open research issues and possible future research directions are discussed.

Key words: entity search; object search; related entity search; similar entity search; knowledge graph

关键字搜索是当前搜索引擎所采用的主流搜索技术,是一种存在性搜索技术,返回给用户包含关键字的网页列表,用户往往需要进一步浏览这些网页并且过滤掉大量无用信息才能找到真正想要的结果.这个过程信息消费代价高,显著降低了用户体验,用户更希望能够直接得到答案.比如,查询“贝拉克·奥巴马的妻子是谁”,用户希望搜索结果是简洁的信息条目“米歇尔·奥巴马”,而不是大量的网页,这种搜索就是实体搜索(entity search).实体搜索的显著特点就是直接给出答案,它关注的是对象,对象可以是各种不同的类别,比如人、电影、公司、小

* 基金项目: 国家自然科学基金(61472426, 61432006)
Foundation item: National Natural Science Foundation of China (61472426, 61432006)
收稿时间: 2016-09-30; 修改时间: 2016-11-23; 采用时间: 2017-01-07; jos 在线出版时间: 2017-01-22
CNKI 网络优先出版: 2017-01-22 16:42:08, http://www.cnki.net/kcms/detail/11.2560.TP.20170122.1642.002.html

说等.例如:查询“贝拉克·奥巴马的妻子”希望得到的就是类别为“人”的具体实体;而查询“汤姆·汉克斯主演的电影”,希望得到的是类别为“电影”的实体列表.文献[1]中通过对用户搜索日志的分析,用户的搜索意图是实体的占所有搜索的比重高达 52%.可见,实体搜索在信息检索中占了非常大的比重.

为了推动实体搜索任务的研究,一些知名的信息检索竞赛也都出现了实体搜索任务.INEX(initiative for evaluation of XML retrieval)从 2007 年开始便有了实体搜索的任务(INEX-XER)[2];TREC(text retrieval Conf.)从 2009 年开始有实体搜索任务[3];QALD(question answering over linked data)[4]是基于链接数据(比如 DBpedia 等)做问答竞赛,其查询结果也是实体或者实体列表;SemSearch(semantic search challenge)关注到实体搜索的重要性,从 2008 年开始举办第一届研讨会及搜索竞赛.另外,SIGIR 从 2011 年开始举办第一届面向实体搜索研讨会(entity-oriented search workshop).国内外各大搜索引擎公司也都推出了实体搜索相关服务,如图 1 所示:当在搜索引擎中输入查询“不爱叫的狗”,搜索引擎直接将狗按喜叫程度进行排序,以图片形式返回给用户,这样的搜索结果一目了然,为用户节省了大量时间,提升了用户搜索体验.类似的搜索还有很多,比如“冬季开花的植物”“省会城市”等.可见,实体搜索越来越引起工业界和学术界的关注和研究.

Fig.1 An example of entity search

图 1 实体搜索示例

本文第 1 节总体概述实体搜索的定义、任务分类及常用的评测指标.第 2 节对相关实体搜索技术在非结构化数据集和结构化数据集上的研究进展进行分类、对比和总结.第 3 节对相似实体搜索技术在两种不同数据集上的研究进展进行综述.第 4 节是对全文的总结及未来研究工作的展望.

1 实体搜索概述

本节首先给出与实体搜索相关的几个术语,然后对实体搜索流程做了简单介绍,同时对实体搜索任务及相关的研究方法做了分类综述.

1.1 相关定义

- 实体(entity):指现实或虚拟世界中具有特定语义的任何对象或者概念都可以看做是实体,用符号 e 表示实体,比如:“贝拉克·奥巴马” “北京” “阿甘正传”等;
- 实体类别(entity class):每个实体都有对应的类别信息,比如实体“北京”是“地方”“城市”.类别体系构成一个层次结构,比如:“机构”是一种类别,机构又包括“公司”“学校”“党派”等不同类别;
- 实体关系(entity relation):表示实体之间的关系,比如“贝拉克·奥巴马”和“米歇尔·奥巴马”两个实体之间的关系为“夫妻”;

- 实体搜索(entity search):指根据实体与给定查询的相关性或相似性对实体做排序,形式化定义为四元组$\{D,q,F,R(q,e_j)\}$,其中,
  (1) D 是数据集.数据集可以是非结构化文档的集合,文档中包含了大量的实体;也可以是结构化的知识库(又称为知识图谱),知识库由大量的实体以及实体之间的关系组成;
  (2) q 是用户查询;
  (3) F 定义为构建数据集表示、查询表示以及它们之间关系的模型框架.例如,对于概率语言模型,该框架是由概率运算和贝叶斯理论组成;
  (4) $R(q,e_j)$是评分函数,输出一个与查询 q 和实体 e_j 有关的实数,用于度量查询 q 和实体 e_j 的相关性或相似性,据此可以对 e_j 进行排序,其中,实体 e_j 存在于数据集 D 中.

1.2 实体搜索流程

实体搜索一般由 3 个步骤组成,基本流程如图 2 所示.首先是对查询做理解分析,包括分析出查询的中心词等;然后在数据集中检索出候选实体;最后再对候选实体做排序,从而得到检索结果.

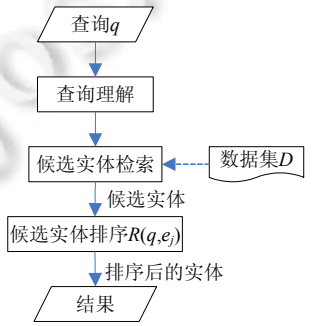


Fig.2 Process of entity search

图 2 实体搜索流程图

在检索开始之前,首先需要对用于支持查询的数据集(非结构化的文档/网页集合,或者是结构化的数据)进行处理.在用户将查询提交到检索系统后,检索系统首先对查询做解析,然后,系统根据对数据集预先建立的索引实现对与查询相关数据的快速检索.如果数据集为非结构化的文档/网页集合,则经过检索系统后,输出相关文档集,然后在相关文档集中做实体检索,从中抽取出候选实体.在将结果提交给用户之前,根据候选实体与查询的相关度或相似度对结果进行排序.

1.3 实体搜索任务分类

实体搜索大致可以分为两类(见表 1).

- 第 1 类是相关实体搜索,输入是各种关键词或者自然语言描述的查询.根据问题答案的实体数目,相关实体搜索又可以分为:单实体搜索,即答案是一个实体,比如“贝拉克·奥巴马的妻子是谁”,此类查询更加关注第 1 个结果的准确率;多实体搜索,答案是多个实体,比如“美国历任总统”,此类查询除了关注准确率之外,还需要关注召回率.单实体搜索和多实体搜索在技术实现上没有太多区分,可将单实体搜索看做是多实体搜索的一个特例;
- 第 2 类是相似实体搜索,输入是几个实体,以这几个实体为例,搜索出与其具有相同语义的其他实体,在 INEX[2],TREC[3]竞赛中又称为实体列表补全或实体集合扩展.这类任务是基于用户已知几个目标答案,比如用户已知几个美国总统,以这几个总统作为例子——“贝拉克·奥巴马、乔治·沃克·布什、威廉·杰斐逊·克林顿”,由系统将其他与之相似的实体补充到列表中,是典型的 Query-by-Example 的实例.

Table 1 Classification of entity search

表 1 实体搜索任务分类

实体搜索任务	查询	查询意图	查询示例	查询示例结果
相关实体搜索	关键词或自然语言描述的查询	单个/多个实体	“贝拉克·奥巴马的妻子”	“米歇尔·奥巴马”
			“美国历任总统”	“贝拉克·奥巴马,乔治·沃克·布什,威廉·杰斐逊·克林顿...”
相似实体搜索	实体或实体集合	多个实体	“贝拉克·奥巴马,乔治·沃克·布什,威廉·杰斐逊·克林顿”	“约翰·肯尼迪,罗纳德·威尔逊·里根...”

1.4 问题空间的划分

虽然实体搜索没有限定数据源,但在实际应用中,根据采用的技术需要预先选定数据源.实体搜索的数据源包括非结构化网页/文档数据和结构化数据,不同数据源在做实体搜索时处理方法也有所不同.本文从实体搜索不同的任务和所使用不同数据源的角度对研究方法做了归类,如图 3、图 4 所示.

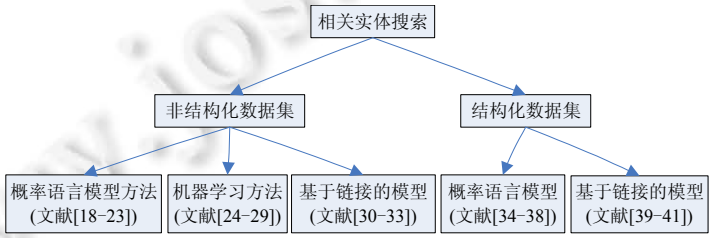


Fig.3 Classification of studies on related entity search

图 3 相关实体搜索研究框图

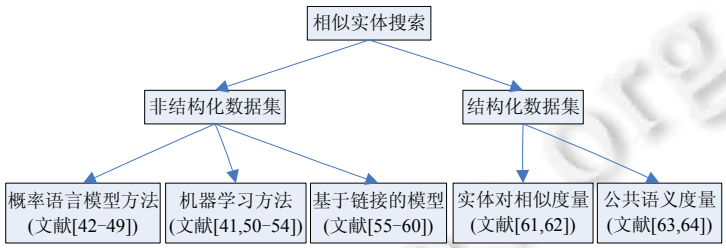


Fig.4 Classification of studies on similar entity search

图 4 相似实体搜索研究框图

1.4.1 数据源概述

实体搜索离不开丰富的数据源.近年来,随着互联网的快速发展,文本、视频、语音等非结构化数据大量涌现.同时,随着开放链接数据(linked open data,简称 LOD)[5]等项目的开展,结构化的语义网知识库数量激增.

(1) 非结构化数据集

目前,已有的工作主要是在网页或者非结构化文本数据集上进行搜索,基本思想都是使用文档作为一个承载实体的桥梁,将查询与候选实体连接起来.分别计算查询与文档的关联性、文档与实体的关联性,或者说是文档刻画实体的程度.文献[6]中主要介绍了两种模型,如图 5 和图 6 所示.其中,第 1 种根据实体出现在不同文本中的上下文,为每个实体创建生成文档(profile),根据这个实体对应的生成文档计算与查询的相关度,对实体生成文档做排序,也就是对实体的排序;第 2 个模型如图 6 所示,根据与查询的相关度,首先对文档进行排序,然后根据实体同文档的相关度再对实体进行排序.

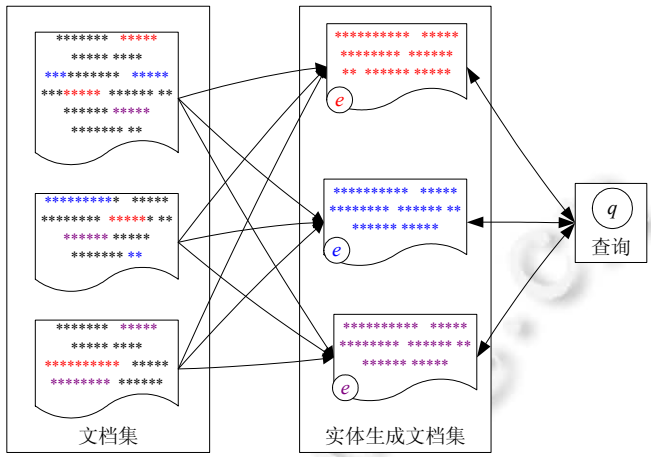


Fig.5 Candidate entity generation model

图 5 基于实体生成文档的模型

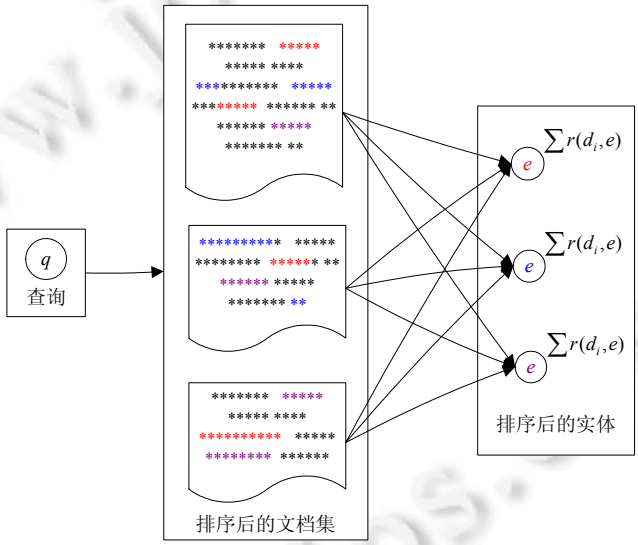


Fig.6 Document generation model

图 6 基于文档的模型

很多非结构化的网页本身就是关于实体的描述信息,比如:Wikipedia 的每个网页都是对一个实体的描述;IMDB 的每个页面也都是对一个电影的介绍;Amazon 的每个页面都是与某个产品相关的信息.所以,利用非结构化网页来做实体搜索是可行的.

(2) 结构化数据集

随着 LOD 等项目的全面展开,语义网数据源的数量激增,大量 RDF(resource description framework)数据被发布.互联网正从仅包含网页和网页之间超链接的文档万维网(Web of document),转变成包含大量描述各种实体和实体之间丰富关系的数据万维网(Web of data).

RDF(资源描述框架)是由 W3C 提出的一种简单、灵活的描述语义网的数据模型,它用来描述实体的信息(属性及对应的属性值)以及实体和实体之间的关系[7].RDF 数据的基本单元是三元组,包括主语、谓语、宾语.其中,宾语也可以作为其他三元组的主语.整个 RDF 知识库可以看做是一个有向并且边上带标记的图(称为

RDF 图),其中,结点表示实体,边表示一个属性或者关系.RDF 图包含丰富实体语义信息,也称为 RDF 知识库.有很多公开的 RDF 知识库,比如 DBpedia[8],Freebase[9]和 YAGO[10]等,每个都包含上百万的实体和数亿的三元组,随着不断地挖掘和自动化的创建,还在不断增长.RDF 知识库为信息检索相关研究比如语义搜索、问答系统[4]提供了新的数据源.2012 年 5 月,Google 发布了知识图谱智能化搜索功能,知识图谱构建所有类型实体,包括人、建筑、城市、国家、电影、小说、艺术品等的关系网络.早期的知识图谱是建构在 Freebase,DBpedia 等公开的 RDF 数据源上,最近,Google 又通过机器学习和数据挖掘的方法,从海量网页中自动发现新的实体和实体间的关系[11].文献[12]中指出,知识图谱将成为下一代搜索引擎的焦点.

结构化数据集上的实体搜索技术主要是利用了数据集的结构化特征,知识是通过实体及其复杂的语义关系来表达,这样的表达方式更便于对知识的深度理解,可以支持将传统搜索引擎中基于关键字的搜索提升到知识理解的层次.

(3) 小结

本节主要对两类数据集做了介绍,其中:基于非结构化数据集的方法主要是利用了同查询语句中词语的共现统计信息来查找实体,而不关注实体之间丰富的语义信息;利用结构化数据集的方法虽然充分利用了结构化数据集上丰富的语义信息,但是相对于网页来说存在知识欠缺问题,包括知识库中缺少新产生的实体,实体之间的关系也没有非结构化数据集上丰富.表 2 列出实体搜索中用到的有代表性的数据集.

Table 2 Representative datasets

表 2 代表性数据集

数据集	特点	规模	数据形式	网址
Wikipedia	基于维基技术的 多语言网络百科全书	51G 的 数据规模	非结构化 网页数据	https://dumps.wikimedia.org/
ClueWeb12	在 2012 年的 2 月和 5 月期间 爬取的网页,被多次用于 TREC 竞赛作为数据集	7 亿 3 千 3 百万 英文网页	非结构化 网页数据	http://www.lemurproject.org/clueweb12.php/
DBpedia	基于 Wikipedia 生成的知识库	600 万实体 95 亿条三元组	三元组	http://wiki.dbpedia.org/
Freebase	开放的、协作创建的 结构化知识库	19 亿条三元组 2300 万实体	三元组	https://developers.google.com/freebase/
Wikidata	与 DBpedia 不同,Wikidata 不仅 提供在线浏览,而且任何人都 可以对词条进行编辑	1500 万实体 4300 万条三元组	三元组	https://www.wikidata.org/
YAGO	基于 GeoNames,WordNet 以及 Wikipedia 开发的知识库	1 亿两千万条三元组 1 千万实体	三元组	http://www.mpi-inf.mpg.de/departments/ databases-and-information-systems/ research/yago-naga/yago/
Probase	由网页抽取构建的概率知识库	5 百万个类别 1255 万实体	三元组	https://www.microsoft.com/en-us/research/ project/probase/

1.5 实体搜索的评测指标

本节综合了信息检索领域中几个竞赛,包括 INEX[2],QALD[4]及 TREC[3]在实体搜索相关任务中用到的评测指标以及实体搜索相关论文中用到的度量指标.这些评测指标除了常用的准确率、召回率以及 F 值之外,还包括:

- Precision@N:在第 N 个位置上的准确率.用户往往只关注前面的几个结果,所以需要在固定的较少数目的结果中计算正确率;
- MRR:对于某些 IR 系统,比如问答系统,只关心第 1 个正确结果的位置,越靠前越好,第 1 个正确结果所在位置的倒数称为 RR.对所有查询的 RR 求平均,得到 MRR;
- MAP:其中,AP 表示平均正确率,对不同召回率上的正确率进行平均.AP 是对单一查询而言,对所有查询的 AP 值进行算术平均,得到 MAP;
- R-Pre:R-Precision 表示检索结果中,在所有相关文档总数位置上的准确率.如某个查询的相关文档总数

为 80,则计算检索结果在前 80 篇文档中的准确率;

- NDCG(normalized discounted cumulative gain):最早在文献[13]中提出了该指标,每个结果不仅仅只有相关和不相关两种情况,而是有相关度级别,比如 0,1,2,3.对于返回结果,认为相关度级别越高的结果越多越好,相关度级别越高的结果越靠前越好.NDCG 能够支持非二值的相关度定义;
- xinfAP:在文献[14]中,xinfAP 被最早提出,使用分层抽样方法来计算平均精度,该指标被应用于实体搜索竞赛 INEX09ER[2]中.

1.6 常用的评测集

本节将对上文中提到的实体搜索相关的几个评测集作介绍,包括评测集的特点、所使用数据集、评测指标并且给出了查询示例,见表 3.

Table 3 Representative testsets

表 3 代表性评测集

评测集	介绍	数据集	评测指标	查询示例
INEX-XER	2009 年 INEX 实体搜索竞赛,返回实体列表	Wikipedia	MAP/xinfAP/P@n	US presidents since 1960
TREC entity	TREC2009 实体搜索任务,关注相关实体搜索	ClueWeb09	NDCG/P@n	Airlines that currently use Boeing 747 planes
SemSearch LS	返回实体列表	BTC-2009	MAP/R-pre/P@n	Axis powers of World War II
QALD	基于链接数据的问答评测	DBpedia	Recall/Precision/F-measure	Who is the major of Berlin?
INEX-LD	INEX 2012 年开始出现基于链接数据的实体搜索,查询使用关键词的形式	Wikipedia	Precision/Recall/AP/MAP/P@n/MRR/NDCG	England football player highest paid

2 相关实体搜索

相关实体搜索的目的是检索出与查询相关的实体.用户的输入一般是包含关键词的自然语言描述的查询语句[2,3,15,16].比如,“1960 年以来美国的总统”,用户的搜索意图是 1960 年以来的美国总统列表,该查询是 INEX[2]中一个查询.查询可能并不符合文法或者语法要求,比如“ben franklin”“american embassy nairobi”等.在竞赛中使用的数据集包括非结构化数据和结构化数据,比如在 INEX Entity Ranking track 的比赛中使用的是 Wikipedia 数据集;在 2009 TREC 实体搜索竞赛中使用的数据集是爬取的网页,也包括 Wikipedia;在 QALD 的竞赛中使用的有结构化的 DBpedia 数据集.本节将分别介绍相关实体搜索在非结构化和结构化数据集上的方法.

2.1 非结构化数据集上的相关实体搜索方法

在非结构化数据集上的相关实体搜索方法主要包括概率语言模型、机器学习方法、基于链接的方法.这些方法都是将文档看做是连接查询与实体的桥梁,在下文中,将依次对各类方法做介绍.

2.1.1 概率语言模型

概率语言模型可以分为 3 类:一类是计算由查询生成候选实体的概率模型,即查询似然模型;另一类是由候选实体生成查询的概率模型,也称为查询生成概率模型或者称为实体似然模型;第三类就是分别将查询与文档建模,然后比较两个模型概率分布的相似性.

(1) 查询似然概率语言模型

先从标准语言模型[17]在文档检索中的应用说起.对文档 d 构建其对应的语言模型 θ_d,检索目标是将文档按照其与查询 q 的相关概率 $P(q|d)$排序,使用贝叶斯公式表示为 $P(d|q)=P(q|d)P(d)/P(q)$,对于每个查询,$P(q)$可以认为是相等的,$P(d)$可以认为是文档的先验概率,与查询是无关的;而 $P(q|d)$可以理解为由某篇文档 d 生成某个查询 q 的概率.文档和查询都是由一个个的词组成的,使用多项式一元语言模型,有:

$$p(q \mid d) = \prod_{t \in q} P(t \mid \theta_d)^{n(t,q)},$$

其中,$n(t,q)$表示词 t 在查询中出现的次数,而文档可以看做是词的多项式分布.

如上介绍了标准语言模型,在实体搜索中,文档集可以直接使用 Wikipedia,每个实体本身就对应一个页面,页面的信息就是描述该实体.也可以在文档中根据每个实体的上下文为实体创建生成文件.文献[18,19]中提到的长文本匹配就是基于查询似然概率语言模型,对于文档的排序结果就是对实体的排序,公式可以写为

$$p(e|q) \propto p(e)p(q|\theta_e) = p(e)\prod_{t\in q} p(t|\theta_e)^{n(t,q)},$$

其中,$p(e)$是实体先验概率;θ_e可以看做是实体的语言模型,是实体在所有词上面的多项式概率分布.

(2) 查询生成概率语言模型

查询生成概率语言模型是计算由候选实体生成查询的概率,又称为话题生成式模型,把查询看作是一个话题,计算由候选实体生成查询的概率 $p(q|e)$.

根据计算概率 $p(q|e)$方式不同,查询生成概率估计又分为两类.

- 一类是候选模型,建立实体 e 的语言模型θ_e,估计由θ_e生成查询 q 的概率 $p(q|\theta_e)=\prod_{t\in q} p(t|\theta_e)^{n(t,q)}$,其中,计算 $p(t|\theta_e)$时,利用文档作为桥梁,首先计算由实体生成文档的概率,然后计算由生成文档生成查询中各个词的概率.另外,在计算时还需要考虑使用平滑技术;
- 另外一类称为文档模型,不对实体构造语言模型,而是直接计算:

$$P(q|e)=\sum_d P(q|d,e)P(d|e),$$

其中,$P(d|e)$计算文档与候选实体之间的关联[20]; $P(q|d,e)=\prod_{t\in q} P(t|d,e)^{n(t,q)}$ 计算文档相关性,即,文档支持候选实体与查询相关的程度,一般基于查询词 t 和候选实体 e 与文档独立的假设[6],公式可写为 $P(t|d,e)=P(t|\theta_d)$.

上述两类生成式概率语言模型是基于候选实体和查询关键词分布是条件独立的,而且查询关键词和候选实体在文档中的语义关系也是忽略的.在文献[21]中,考虑了查询词与候选实体之间的关联,使用两者在文档中的距离来度量其关联程度.

(3) 扩展的概率语言模型

如上介绍的都是单方向的通过文档语言模型生成查询的概率,或者考虑查询语言模型生成文档的概率.另外一种做法不是从单方向来直接生成对象,而是查询和文档同时生成语言模型,然后比较这两个模型之间的差别.文献[22]对这 3 种方法做了比较,如图 7 所示:图 7(a)为查询似然,图 7(b)为查询生成语言模型,图 7(c)是利用模型比较方法.Balog 在文献[23]中也利用了图 7(c)的方法,对于给定查询 q,计算查询模型和文档模型的 KL 距离,同时也利用了实体类别信息.查询和实体都使用基于关键词的概率分布和基于类别概率分布来模拟,每个实体表示为在关键词和类别上的概率分布对: $e=(\theta_e^T,\theta_e^C)$,其中, θ_e^T 是关键词概率分布, θ_e^C 是该实体关于类别的概率分布.同样,查询也表示为两个概率分布对: $q=(\theta_q^T,\theta_q^C)$.对于实体的排序,就是基于实体与查询的概率分布相似度来模拟.该相似度通过两个概率分布的 KL 距离来实现,KL 距离越大,两者的分布越不相似,他们之间的相关度越低;反之越高.

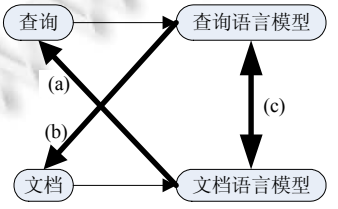


Fig.7 Probabilistic language modelings

图 7 概率语言模型建模的 3 种方式

(4) 模型比较

生成式模型是从统计的角度来表示数据分布,充分利用先验知识.Lafferty 和 Zhai 在文献[22]中给出的实验结果表明,基于模型比较的方法比查询似然和查询生成概率估计方法都好.

2.1.2 机器学习模型

近年来,Learning to Rank(L2R)被广泛应用于文档排序,Liu 在文献[25]中给出了关于 L2R 的综述文章.而实体搜索关键也是排序问题,所以 L2R 的方法也可以应用.Sorg 和 Cimiano 在文献[26]中将实体搜索看做是二分类问题,通过使用多层感知机和逻辑回归模型作为分类器.通过选用一些精心设计的特征,实验表明:在 Yahoo Answers 抽取的数据集上优于生成式语言模型.Yang 等人在文献[27]中将排序的 SVM 模型应用于 ArnetMiner 的专家排序.LADS 系统[28]在 TREC 的相关实体查找任务中,基于 L2R 的方法对候选实体做排序. LADS 包括 4 个主要部分:文档排序、实体抽取、特征抽取及实体排序.

文献[29]主要关注的是实体排序,预先定义一些不同层次的特征,比如词的特征,有 TF、IDF 值;文章的特征,如文章的长度等;出现位置的特征等.提出使用线性组合模型将不同的特征组合,应用有监督学习方法来对实体做排序.在专家搜索任务上,指标 MAP 可以达到 0.268.

Macdonald 和 Ounis 提出一种有监督集成学习方法[30],将不同的学习器整合在一起,包括文档权重模型(比如 TFIDF 以及 BM25)、投票模型等,把各个学习器的结果进行整合,从而得到比单个学习器要更好的效果.

2.1.3 基于链接的模型

基于链接的模型最早是应用于网页检索,其中,PageRank[31]和 HITS[32]是最著名并且应用最广的两个算法.概率语言模型也可以看做是由文档和候选实体组成的图,文档和实体之间存在连边,得到文档排序后,将相关度传播给实体.在相关实体搜索中,利用基于链接的方法,将文档和实体看做是图中的结点,通过他们直接的链接关系构造图结构,其中,文档与文档、文档与实体、实体与实体都有连边.

实体搜索可以看做是这样一个过程:随机选择一个文档或者候选实体;待浏览文档结束后,可能会跳转到文档中涉及到的某个实体或者与这个文档相关的其他文档;待浏览实体结束后,跳转到提及该实体的其他文档或者与这个实体相关的其他实体.这个过程与随机游走很类似.Serdyukov 等人在文献[33]中提出了针对特定领域的相关性概率多步传播模型.基于链接的实体搜索方法有:(1) 无限随机游走模型,把实体搜索看做是一个不会停止的过程,如同用户的信息需求也是一个不会终止的过程;(2) 另外一种模型是会被吸收的随机游走,游走到某个候选实体结点就会停止.文献[33]中指出:两种方法都要优于一步的相关性传播方法,也就是概率语言模型;另外,无限随机游走模型稍微好于会被吸收的随机游走模型.文献[34]中介绍了一种应用于专家检索的由文档到相关候选实体的多步传播算法,利用可吸收的随机游走模型构造文档与候选实体图.利用 TREC 2005-2007 的数据集,在 W3C 2006 数据集上,MAP 可以达到 0.398.

2.2 结构化数据集上的相关实体搜索方法

在结构化数据集上,相关实体搜索方法包括概率语言模型方法和基于链接的模型..

2.2.1 基于概率语言的模型

在结构化数据集上使用概率语言模型,首先要为每个实体创建生成文档.文献[35–39]中,将与每个实体相关的三元组看做是实体的生成文档,计算实体的生成文档与查询的相关度,然后排序,便得到了候选实体的排序.

以文献[37]为例,详细介绍概率语言模型在结构化数据集上的应用.首先是实体生成文档如何创建问题,提到了 3 种方法,我们结合图 8 来对结构化数据集上实体生成文档做介绍.

图 8 是关于实体“贝拉克·奥巴马”的知识片段.

- 一种是只使用与该实体存在联系的实体,不关心他们之间的关系(谓词)类型,称为无结构的实体模型,此种模型的生成文档见表 4;
- 另外一类是将谓词分为 4 类:Name(表示实体的名称)、Attributes(表示实体具有的属性值,即实体通过该谓词对应的宾语为值,而不是实体)、OutRelations(表示实体出边指向的实体)、InRelations(表示实体入边指向的实体),此类称为结构化实体模型,此种模型的生成文档见表 5.

在 2010 年和 2011 年的 SemSearch 竞赛测试集,基于 Billion Triple Challenge2009 数据集,非结构实体模型 2010 年的 MAP 最高是 0.212 5,NDCG 为 0.384 8;2011 年 MAP 最高为 0.207 2,NDCG 为 0.294 7;而结构实体模型在 2010 年 MAP 高达 0.281 6,NDCG 为 0.494 3;2011 年 MAP 可以达到 0.261 4,NDCG 可以达到 0.396 8.论文还分析到:异构信息网络中存在大量的谓词,谓词的权重是各不相同的;对于一个实体而言,一般只有几个不同的谓词;结构实体模型中可以明显看出是有效果的,但还是没有利用到他们之间的语义信息,在文献[36]中提出了新的层次实体模型:第 1 层还是实体的 4 个类别(Name,Attributes,InRelations, OutRelations),第 2 层是各个不同的谓词,此种模型的生成文档见表 5.

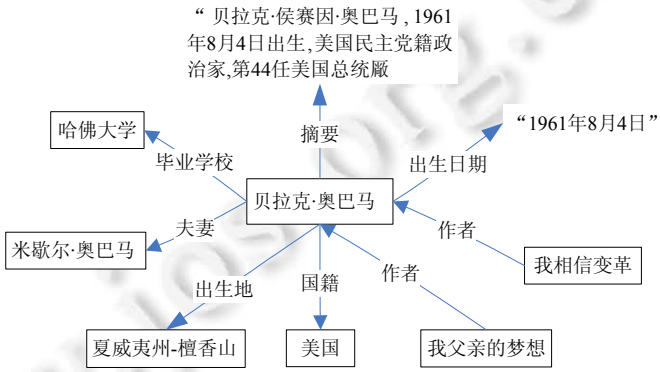


Fig.8 Knowledge fragment of Barack Obama

图 8 贝拉克·奥巴马在知识库中的知识片段

Table 4 An example of unstructured entity model

表 4 无结构的实体模型示例

实体生成文档
贝拉克·奥巴马 米歇尔·奥巴马 贝拉克·侯赛因·奥巴马(Barack Hussein Obama),1961 年 8 月 4 日出生,美国民主党籍政治家,第 44 任美国总统... 1961 年 8 月 4 日 美国 夏威夷州-檀香山 哈佛大学 我相信变革 我父亲的梦想

Table 5 An example of structured entity model

表 5 结构和层次实体模型示例

结构的实体模型		层次实体模型		
谓词类型	值	谓词类型	值	
			谓词	谓词对应的值
Name	贝拉克·奥巴马	Name	名字	贝拉克·奥巴马
属性	贝拉克·侯赛因·奥巴马(Barack Hussein Obama),1961 年 8 月 4 日出生,美国民主党籍政治家,第 44 任美国总统... 1961 年 8 月 4 日	属性	摘要	贝拉克·侯赛因·奥巴马(Barack Hussein Obama),1961 年 8 月 4 日出生,美国民主党籍政治家,第 44 任美国总统...
			出生日期	1961 年 8 月 4 日
OutRelations	美国 夏威夷州-檀香山 哈佛大学	OutRelations	国籍	美国
			出生地	夏威夷州-檀香山
			毕业学校	哈佛大学
InRelations	我相信变革 我父亲的梦想	InRelations	作者	我相信变革
			作者	我父亲的梦想

3 种不同模型使用图 9 的形式来表示((左)无结构实体模型,(中)结构实体模型,(右)层次实体模型).文献[35]

中通过实验表明,层次实体模型比非结构化实体模型以及结构化实体模型效果都要好.3 种模型的主要区别在于对于谓词或关系的处理方式不同,其中:无结构的实体模型不考虑谓词的类型及具体谓词;而结构化实体模型将谓词分为 4 种类型;层次模型不仅考虑谓词的分类,还包括具体谓词的值.在具体应用时,考虑实体间关系的模型效果更好.

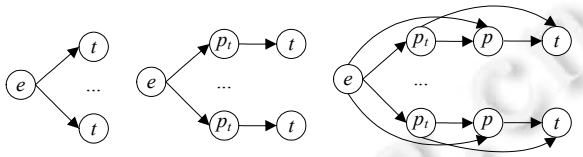


Fig.9 Diagrams of three different types of entity model

图 9 3 种不同的实体模型图示

2.2.2 基于链接的模型

结构化数据集具有丰富的语义关系,而经典的基于链接的算法比如 Pagerank,SimRank 等都不关心边上的语义信息,所以简单地使用 Pagerank,SimRank 等方法不能体现出结构化数据集的优势.文献[40]提出根据知识库的本体信息,对于类别之间的链接关系做排序,但是不涉及具体实体.文献[41]考虑了不同实体之间不同联系具有不同权重,通过领域专家来学习权重,然后对目标实体进行排序.文献[42]对于不同关系权重的获取不再仅仅依赖于专家标注,而是在专家标注部分权重的基础上,再通过模拟退火算法来学习到更多权重.

2.3 相关实体搜索方法总结

在非结构化数据集上相关实体搜索的方法主要是概率语言模型,文献[22]对这几类方法分别做了实验,结论是:基于模型比较的方法效果最好,但是此方法本身计算代价要比其他两类高.在结构化数据集上,概率语言模型分别展示了 3 种不同的实体建模方法,其中,层次化模型效果最好.因为层次化不仅考虑了属性信息,还对不同的属性分配不同的权重,对于实体搜索有更大帮助.

表 6 是基于非结构化数据集 Wikipedia 在 INEX07,INEX08 上的精度比较.几类方法都是基于概率语言模型,其中,BBR[23]算法还使用了反馈技术;KK[18]算法利用了伪相关反馈技术对查询进行扩展,同时还利用了文档之间的链接关系,将文档的得分进行传播;CGSDW[19]算法利用了重排序技术.目标实体类别对于实体搜索是非常有意义的,表 6 对如上提到的 3 类方法分别比较了是否考虑目标实体类别对精度的影响,其中,CC_BBR,CC_KK,CC_CGSDW 考虑了目标实体的类别.可以看出,考虑目标实体类别对于精度有很大程度提升.

Table 6 Comparisons of results on the INEX07, INEX08 topic sets

表 6 在 INEX07,INEX08 上的方法对比

Models	INEX07	INEX08
	MAP	xinfAP
BBR[23]	0.180	0.135
CC_BBR[23]	0.255	0.312
KK[18]	0.184	0.159
CC_KK[18]	0.262	**0.352**
CGSDW[19]	0.075	0.089
CC_ CGSDW[19]	**0.265**	0.334

目标实体类别匹配技术主要包括:

1) 精确匹配,要求候选实体类别与查询目标实体类型完全一致.比如查询“克里斯蒂安·贝尔参演的电影”,查询要求目标实体类型是“电影”,如果“金陵十三钗”的类型是“中国电影”,那即使“金陵十三钗”是正确的答案,也会因为类别精确匹配影响正确的结果.文献[23]是精确匹配;
2) 模糊匹配,不要求目标实体类别与查询目标实体类别精确匹配,他们之间的相关性通过文本相似性来计算.例如查询目标实体类别是“美国电影”,而实体类别是“中国电影”,由于两种类别之间具有相关

性,所以实体还是会被赋予一定分值.文献[18]利用了类别的模糊匹配;

3) 结构化匹配,利用实体类别同查询目标实体类别的上下位关系.上例查询中,查询目标实体类别是“电影”包含“中国电影”这个类别,所以“金陵十三钗”也是正确的结果.文献[19]就是利用了类别的结构化匹配方法.

可以看出:目标实体类别精确匹配要求过于严格,没有考虑类别之间的上下位关系,会影响查询结果精度.另外,基于候选实体生成的两种算法 BBR 和 KK 在不考虑类别匹配的情况下,精度都要高于采用文档模型的 CGSDW 算法.

3 相似实体搜索

相似实体搜索输入的查询 q 是实体集合(这些实体也被称为种子实体),通过搜索引擎找到与给出的实体具有相同语义的其他实体.相似实体搜索又称为实体集合扩展,在竞赛 INEX[2]又称为列表补全(list completion),文献[42]中称为 QBE(query by example).

由于用户背景知识所限,有时对于一个查询不能给出很好的定义,从而通过传统的搜索引擎得不到期望的结果.设想:在欣赏一个画展时,用户观察到高更和梵高的画风很相似,用户想搜索与这两个画家具有相似画风的其他画家,然而用户并不了解他们属于什么画风.在搜索引擎中输入查询“与梵高、高更画风相似的画家”,返回结果如图 10 所示,可以看出,目前的搜索结果并不能满足用户需求.从返回的页面里面没有找到与梵高、高更相似的画家,虽然用户不清楚他们的画风是什么,但是可以把梵高、高更作为例子(种子),通过相似实体搜索找出与他们相似的画家并且给出解释(比如,他们都是“后印象主义”).因为给出对他们相似性的解释后,用户还可以用来做查询扩展,把查询修订为“后印象主义的画家 梵高 高更”.实体集合扩展最早是 Google Set 提出,目前应用于 Google Driver,Google Doc 中.另外,还可以应用到推荐系统中,比如:可以根据用户的点击记录,挖掘出背后的语义,可能是来自同一个作者的几本书,或者是同一种类型的书等等,从而做出更准确的推荐.

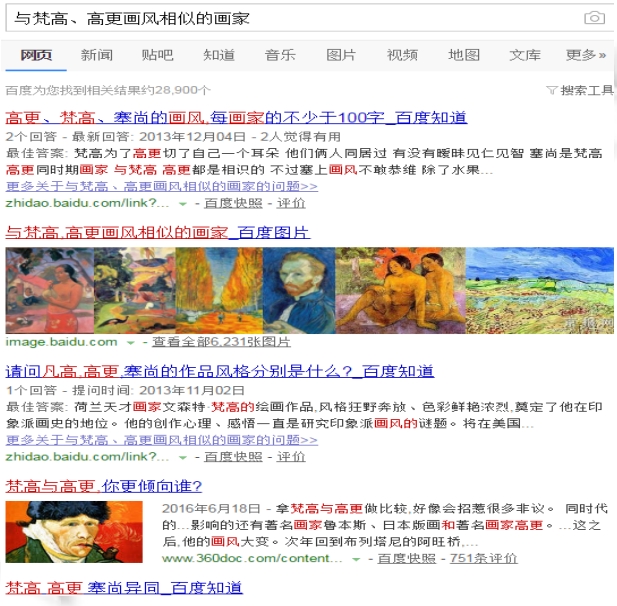


Fig.10 An example of similar entity search

图 10 相似实体搜索示例

相似实体搜索虽然绕过了自然语言处理的难题,但是如何通过给出的种子实体找到与之相似的其他候选实体也是很有挑战的,包括:

1) 如何找到种子实体之间的共同语义特征,如何定义语义特征.例子实体之间共同的语义特征可能是间接的语义,比如给出“巩俐、章子怡、周冬雨”作为例子,用户的搜索意图是找出“谋女郎”,给出的种子与“张艺谋”之间的关系是一种间接关系,因为这几个种子实体参演了某几部由“张艺谋”导演的电影,他们通过谓词“参演”和“导演”与“张艺谋”建立联系,是一种间接关系.另外,在找共同语义特征时,还需要考虑数据集中存在知识缺失;
2) 找到共同语义特征后,如何设计排序模型,并对候选实体做排序.有效的排序模型才能保证得到高的精度;
3) 在大数据的背景下,做到支持在线查询,快速地找到与给定的种子集合相似的实体,也是非常具有挑战的.

本节将分别在非结构化数据集和结构化数据集上介绍几种典型的相似实体搜索方法.

3.1 非结构化数据集上的相似实体搜索方法

相似实体搜索在非结构化数据集上的方法主要包括 3 种:概率语言模型、机器学习方法、基于链接的方法、.

3.1.1 概率语言模型

基于概率语言模型的相似实体搜索方法利用了生成式语言模型,计算出与给定的种子实体相似性最高的候选实体[43−50].文献[46]最早提出基于种子实体与候选实体具有相同概率分布的假设,利用贝叶斯公式来发现其他候选实体,同时也要考虑候选实体在整个文档中的出现概率.为了提升计算性能,在论文中用到了共轭先验分布,从实验结果看,准确率有很大提升.其他工作[44,47,48]都是基于文献[46]的基础上做的扩展.

3.1.2 机器学习方法

文献[51]提出了半监督的不带标签的正例(PU)学习算法来获得候选实体,该算法是在贝叶斯分类器和最大期望算法基础上使用了间谍技术.算法思路是:首先在正例集合(可以认为是种子集合)中抽样出部分实体,标记为 SP,称为间谍;然后以正例与 *SP* 的差集作为学习的正例,未标记的实体集合与 *SP* 的并集作为负例来构造贝叶斯分类器;最终,针对所有实体都可以学习到一个属于正例的概率值,根据该数值也就得到了候选实体排序.Lim 在文献[42]中提出利用种子实体,通过机器学习方法挖掘出查询的潜在语义及相关本体信息,最后执行查询并且生成查询结果.Sadamitsu 在文献[52,53]中提出使用判别式模型改进的 bootstrapping 方法.Jindal 在文献[54]中需要在种子节点集合不仅仅给出正例,还需要给出负例,通过在文本集上的学习得到基于推理的方法来实现相似实体查找.Bing 等人在文献[55]中提出利用条件随机场模型的半监督学习方法在网页中根据给出的种子实体找出新的相似实体.

然而,通过机器学习方法做实体搜索存在数据稀疏问题,对于信息量少的长尾词搜索结果不好;另一个问题是,由于网页或者知识都是动态变化的,某段时间用来做训练的数据集不一定能够很好地适应未来的应用.

3.1.3 基于链接的模型

在非结构化数据集上查找与给定的种子实体集合相似的实体,具有代表性的工作是 CMU 开发的 SEAL 系统[56−60].SEAL 的思想是,认为包含了种子实体的网页含有其他候选实体的可能性也很大.SEAL 系统包括 3 种模块,分别是抓取模块,使用所有例子实体,构造一个查询,利用 Google API 检索前 N 个结果的 URLs,将爬取的网页文档提交给抽取模块;抽取模块,利用网页的半结构化特征,从抓取下来的网页中抽取或者说学习模板,然后根据学习到的模板找出候选实体;排序模块,按照与种子的相似程度进行排序,使用种子节点、候选实体节点、模板以及网页构造成一个图,使用随机游走算法获取排序后的候选实体列表.

SEAL 系统是与语言无关的(种子支持各种语言,中文、英文、日文等等;处理文档的类型可以是 HTML 或 XML),并且不需要预先标注训练数据,直接使用互联网语料,模板是自动学习获得的.但由于需要根据种子节点信息实时抓取网页,所以会有较多的时间开销.另外,因为网页的动态性,不同时期的查询结果可能不相同.SEAL 系统不能给出种子实体具有的语义特征,对于查找结果的解释性差.

另外一个工作 SEISA[61]使用网页中的表数据和查询日志先离线构造二部图,提出了两个实体排序指标,既要保证候选实体同种子实体之间的相似性高,同时还要保证候选实体之间具有较高的相似性.在实验部分给出

4 种类型的种子实体,分别是国家、颜色、相机、床垫,与随机游走及 Google Sets 和 SEAL 做了比较.整体而言,SEISA 在准确率和召回率上效果都要更好.然而该方法受限于数据源,如果使用的表数据或者查询日志中没有出现某个种子实体,将会影响查找结果的精度.

3.2 结构化数据集上的相似实体搜索方法

在结构化数据集上的相似实体搜索可以充分利用数据集结构化和语义丰富的特征,主要包括两类方法:第 1 类就是基于实体对相似性度量;第 2 类是通过找出种子实体之间的共同语义,然后来检索相似实体.

基于语义度量的方法就是在结构化数据集或者知识库上度量语义,从而得到与种子实体相似的候选实体.基于链接的模型是基于实体之间的链接关系,此类方法认为:如果两个实体与同一个实体相连,那么这两个实体是相似的.本节将对这两种方法做综述介绍.

3.2.1 实体对相似度量

在结构化数据集上搜索相似实体最具有代表性的工作是由 Sun 等人提出的 PathSim[62]方法,PathSim 基于语义计算实体相似,以 DBLP 数据集的数据模式为例,如图 11 所示,该数据集主要包括 4 种类型的结点:论文、作者、会议、关键字.不同的语义在查找相似实体时侧重也各不相同,比如:“会议-论文-作者-论文-会议”侧重于查找具有相同作者的会议,而“会议-论文-关键字-论文-会议”侧重于查找具有相同关键字的会议.基于此,作者提出了基于不同语义(路径)定义两个实体相似性的度量方法.文章中通过实验表明:PathSim 在 DBLP 数据集上准确率可以达到 0.745,显著高于随机游走、个性化 PageRank、SimRank 方法.

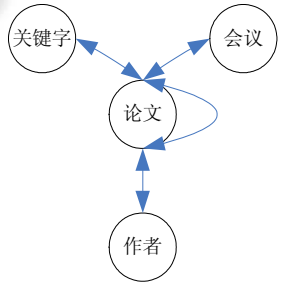


Fig.11 Data schema of DBLP

图 11 DBLP 数据模式

LDSD(linked data semantic distance)[63]基于 RDF 知识库的结构化来度量实体之间的语义距离,在计算两个实体之间语义距离的时候,既考虑了两个实体经过某个谓词直接连接的情况,也考虑了两个实体经过相同的谓词都与某个中间结点相连的情况(也就是两个实体之间存在间接关联关系).文献[63]中将该方法应用于音乐领域的推荐.

这些实体对相似度量都采用度量两个实体之间的相似性,而不是一个实体同一个实体集合的相似性.可能存在某个候选实体与某个种子实体在某些特征上很相似,但是与种子集合中实体的共同特征并不相似.所以,这种基于实体对相似的度量并不适用于相似实体搜索.

3.2.2 基于种子集合共同语义的方法

此类方法是基于分析种子集合中所有种子共同语义特征,然后根据共同特征再搜索满足这些特征的其他实体.

由 Metzger 等人提出的 QBEES[64]是根据知识库检索出所有种子都满足的共同语义,该方法存在两个问题:(1) 没有考虑知识库的知识不完备性,如果种子集合中有某个种子由于知识缺失不符合某个语义特征,而这条语义特征正是用户期望的查询语义,就会影响查询结果的准确性;(2) 该方法没有考虑间接语义,比如,上文中提到的例子种子集合为“巩俐、章子怡、周冬雨”,用户的查询意图是“谋女郎”,这几个种子与张艺谋之间是一种间

接的关系,通过“参演”和“导演”两个谓词连接的,而 QBEES 方法没有考虑这种间接语义.

实体之间的相似性也可以通过使用关联规则挖掘的方法 ARM[65]来计算,具有相同谓词的实体可以认为在某种程度上具有相似性,具有相同谓词-宾语对的实体也可以看做是相似实体.将谓词或者谓词-宾语的组合看作是属性,然后,具有该谓词或者谓词-宾语属性的主语看做是值,构造出“属性-值”对的集合,也就是项集.在进行实体集合扩展时,计算同种子节点具有最高支持度的其他候选实体,然后按照支持度得分的高低排序.

知识库中的关系在实现相似实体搜索中是非常重要的,关系用来限定语义,如果不考虑关系,将会影响到结果的精度.比如“人工智能 林肯 辛德勒名单”,在知识库中,3 个种子实体与“史蒂文·斯皮尔伯格”都存在关系,3 部电影的导演都是“史蒂文·斯皮尔伯格”;另外,“人工智能”的编剧也是“史蒂文·斯皮尔伯格”,如果不考虑具体的关系限定,在返回结果中会将与“史蒂文·斯皮尔伯格”有关的电影都返回,既包括由“史蒂文·斯皮尔伯格”所导演的,也包括“史蒂文·斯皮尔伯格”编剧的,与用户的查询意图不符,影响结果的精度.

3.3 相似实体搜索方法总结

为了进一步比较说明各种方法的不同,本节使用一个示例来将上文中提到的在非结构化及结构化数据集上的典型算法对于相似实体查找结果做展示和分析.

查询示例给出的种子实体集合为“Apollo_13,Philadelphia,Forrest_Gump”,用户的查询意图是查找“汤姆汉克斯参演的电影”.表 7 展示了 4 种典型算法的 Top-10 结果(该查询是 INEX[2]中的一个查询,标准答案由 INEX 提供,结果实体中带有下划线的不在给出的标准答案集合中,视其为错误结果),其中,

- SEAL 是基于非结构化数据集上的检索结果,是基于统计与种子实体的共现来对候选实体排序.3 个种子实体与实体“Tom_Hanks”共现频繁,所以“Tom_Hanks”也出现在结果列表中;另外,SEAL 方法没有用到种子实体的共同语义信息;
- HeteSim 方法是基于实体对相似度量,结果中,实体“Contact”与“Forrest_Gump”的相似性很高(他们具有相同的导演、相同的制片等共同语义),但是 Contact 与种子实体集合并不相似(因为“Tom_Hanks”并没有出演“Contact”);
- QBEES 方法的准确率最低,其原因是知识库存在知识缺失,“Apollo_13”与“Tom_Hanks”之间不存在参演的关系,这种基于所有种子实体都需要满足的共同语义的方法就不能利用这个与用户查询意图匹配的语义特征,只能利用到这些种子实体都满足的一些很宽泛化的特征(比如类型都是电影、都是美国电影)进行候选实体的扩展;
- ARM 方法有一定的容错性,可以在知识库中找到满足用户意图的特征,但是由于对语义特征的排序模型不够有效,所以在 Top-10 的结果中还是有错误实体.

Table 7 A case study of similar entity search

表 7 相似实体搜索代表方法结果示例

SEAL	HeteSim	QBEES	ARM
Saving_Private_Ryan	The_Polar_Express_(film)	Remember_Me_(2010_film)	Cast_Away
The_Green_Mile	Your've_Got_Mail	We_Are_Marshall	Saving_Private_Ryan
Cloud_Atlas	Contact	Mommie_Dearest_(film)	Cloud_Atlas
Captain_Phillips	Cast_Away	Dances_with_Wolves	Sleepless_in_Seattle
Sleepless_in_Seattle	Death_Becomes_Her	Fame_(1980_film)	Dragnet_(1987_film)
Your've_Got_Mail	Back_to_the_Future_Part_III	101_Dalmatians_(1996_film)	Big_(film)
A_League_of_Their_Own	Back_to_the_Future_Part_II	The_Princess_Bride_(film)	The_Da_Vinci_Code_(film)
Tom_Hanks	Flight_(2012_film)	Smokey_and_the_Bandit	Next_(2007_film)
Radio_Flyer	The_Silence_of_the_Lambs_(film)	The_Commitments_(film)	Oliver_the_Eighth
Rita_Wilson	Melvin_and_Howard	The_Good_Shepherd_(film)	Rhythm_in_a_Riff

通过对上文的分析,我们再对各种方法分别从所使用的数据源、对结果的解释性、是否考虑知识库缺失这 3 个角度做总结比较,见表 8.其中,解释性是指该方法是否能够对相似实体搜索结果给出语义解释;知识库是存在知识缺失的,如果不考虑知识不完备,对搜索结果会有一定的影响.

使用 INEX09 测试集和 QALD(由 QALD-2,QALD-3 及 QALD-4 去掉重复查询后构成)测试集,在每个查询

对应的答案中的实体列表中抽取出不同数目的实体作为种子,测试集的说明见表 8.各种不同方法在不同种子数目时实验结果见表 9,其中,结构化数据集使用的是 DBpedia 3.9 版本.

Table 8 Comparison of similar entity search methods

表 8 相似实体搜索方法比较

方法名称	数据源类型	解释性	是否考虑知识库缺失
SEAL	非结构化	N	–
SEISA	非结构化	N	–
HeteSim	结构化	Y	N
QBEES	结构化	Y	N
ARM	结构化	Y	Y

Table 9 Characteristics of the datasets

表 9 测试集描述信息

	INEX	QALD
查询数目	52	60
具有间接语义的查询数目	4	5
混合测试集中具有 2 个种子的查询数目	7	17
混合测试集中具有 3 个种子的查询数目	33	18
混合测试集中具有 4 个种子的查询数目	8	17
混合测试集中具有 5 个种子的查询数目	4	8
查询对应答案包含的的平均实体数目	31	42

基于表 10 中的数据,我们发现:LDSD 和 HeteSim 在两个数据集上的性能都比较差,说明基于实体对相似度量来发现相似实体是不够的.

Table 10 Comparisons of results on the INEX and QALD

表 10 在 INEX,QALD 上的方法对比

方法	种子数目	INEX			QALD		
		p@10	MRR	R-pre	p@10	MRR	R-pre
SEAL	2	**.412**	.542	**.327**	.377	.550	.269
LDSD	2	.200	.461	.166	.122	.264	.113
HeteSim	2	.133	.267	.114	.305	.508	.242
QBEES	2	.338	**.556**	.282	.400	.654	.369
ARM	2	.287	.496	.244	**.422**	**.662**	**.377**
SEAL	3	**.433**	**.547**	**.377**	.363	.591	.340
LDSD	3	.210	.401	.184	.243	.270	.131
HeteSim	3	.158	.309	.129	.312	.557	.279
QBEES	3	.317	.532	.256	.440	**.688**	.423
ARM	3	.260	.435	.224	**.468**	.665	**.446**
SEAL	4	**.383**	**.530**	**.339**	.350	.539	.354
LDSD	4	.235	.408	.179	.163	.308	.153
HeteSim	4	.146	.281	.120	.287	.532	.271
QBEES	4	.312	.451	.229	**.453**	**.668**	**.452**
ARM	4	.256	.493	.222	.430	.716	.420
SEAL	5	**.340**	.418	**.311**	.317	.535	.352
LDSD	5	.292	**.535**	.208	.145	.282	.153
HeteSim	5	.158	.300	.133	.283	.507	.273
QBEES	5	.215	.342	.169	**.428**	.638	**.449**
ARM	5	.267	.484	.239	.418	**.665**	.426
SEAL	MIX	**.397**	**.644**	**.330**	.347	.592	.335
LDSD	MIX	.227	.496	.204	.155	.292	.141
HeteSim	MIX	.169	.339	.136	.287	.510	.263
QBEES	MIX	.371	.591	.299	.408	.626	.412
ARM	MIX	.394	.535	.273	**.443**	**.646**	**.433**

在有 3 个种子的 INEX 测试集上,SEAL 的 R-pre 最高,这是因为 SEAL 使用 Google 搜索引擎来检索包含种子的页面,我们发现:在返回的页面中有 INEX 测试集的答案文件,对结果的提升有很大帮助.在 QALD 测试集

上,ARM 和 QBEES 性能较好,它们都使用是基于种子集合共同语义的方法.另外,几种方法在 QALD 测试集上的表现一般都优于 INEX,其中一个原因是因为 QALD 本身就是基于链接数据设计的查询,查询语句可以直接转成 SPARQL 查询,而 INEX 中很多查询与 DBpedia 知识库中的实体或者谓词都不能对应,从而影响了放在 INEX 测试集上的性能.另外,在只有 2 个种子时,性能都很差,这是因为种子数太少不利于找到种子之间的共同语义.当种子数从 2 增加到 5 时,SEAL 方法的性能有提升,只是提升的显著性随着种子数大于 3 之后就不再明显.对于 QBEES 来说,随着种子数的增长,性能反而降低,尤其是在 INEX 上.这是因为 QBEES 需要所有的种子都满足某个共同语义才可以,如果种子数增加,考虑到知识库中存在知识缺失,找到所有种子都满足的共同语义难度增加.

4 总结及未来研究方向

本文从实体搜索的任务以及实体搜索所使用的数据源这两个维度,对目前已有的实体搜索方法进行了分析和对比.目前的研究工作虽然取得了一些成果,但有些问题仍然值得深入探讨研究.

1) 数据融合

在第 1.4.1 节中分析了非结构化数据与结构化数据的特点,其中:结构化数据质量高、处理相对简单,但是不能及时地包含新出现的实体和新的关系,而且存在知识欠缺问题;而非结构化数据尤其是网页更新很快,数据量要远大于结构化数据.将非结构化数据和结构化的知识库相互融合,两者可以相得益彰.非结构化数据中,实体、关系更为丰富,可以用来补充到结构化数据中.比如,知识库中有三元组(冯小刚,配偶,徐帆),使用关系“配偶”来表示“冯小刚”和“徐帆”两个实体的关系.如果两个在知识库中具有某种关系的实体在相同的句子或者文章中出现,那么他们在句子或者文章中描述的关系很可能与知识库中的关系相同.基于此,观察从网页中抽取出的包含实体“冯小刚”和“徐帆”的句子有:“1999 年,徐帆与导演冯小刚结婚”“冯小刚发表长微博,深情表白妻子徐帆”“冯小刚和现任老婆徐帆为什么不生子内幕”,等等,这样就可以挖掘出知识库中的关系“配偶”同文档中挖掘出来的“结婚”“妻子”“老婆”是相同的,从而可以扩充知识图谱中的关系.这方面的研究在文献[66,67]中有详细阐述,本文不再赘述.另外,使用结构化来表达知识便于用户理解使用,同时可以用来扩展查询.比如,用户想检索“中国研究语义网的实验室”,在很多搜索引擎的搜索框中输入关键词“实验室 语义网 中国”,检索出的结果几乎没有符合用户预期的;而如果能借用结构化的知识,可以发现,“实验室”与“语义网”之间的关系是“研究领域”“实验室”与“中国”之间的关系是“所在国家”.另外,基于结构化的知识库还可以挖掘出关系之间的推理.比如,如果要知道某个“实验室”的研究领域,可以通过实验室所发表论文的领域获得.因此,基于结构化的关系推理可以用来扩充查询,从而可以得到更好的查询结果.文献[68]就是基于融合的数据集做实体搜索,这方面研究还处于初步阶段,缺乏通用的模型.

2) 搜索结果可解释性

在表 8 中对相似实体搜索方法比较时,将解释性也作为一个维度.对结果的解释性,是为用户提供答案的“证据”.结果的可解释性是非常有必要的,因为它搭建了“领域专家”与用户之间的连接,尤其是在一些需要很深厚的专业知识的领域,比如医学诊断或者法律判决等,模型的可解释性可以帮助用户更好的理解结果并且提升结果的可信度.近两年,在推荐系统[69,70]、社区发现[71]中,对于结果可解释性(“证据”)都有所研究.在基于非结构化数据集的实体搜索中,对结果解释性的研究并不多,基于结构化数据集的相似实体搜索算法,比如 ARM,QBEES 都是基于给定的种子实体具有的共同语义挖掘候选实体,而种子具有的共同语义可以看做是一种解释.然而这两种方法本身也存在问题,其中,QBEES 算法没有考虑知识库中的知识缺失,需要所有的种子都满足共同语义才可以,另外,该算法没有考虑间接语义的情况,这些都导致挖掘出来的共同语义有缺失.另外,ARM 算法虽然考虑了知识缺失和间接语义的情况,但是对于挖掘出来的共同语义缺乏有效的排序算法.

3) 数据质量问题

数据质量对于搜索结果具有直接的影响,使用高质量的数据源是得到有效搜索结果的保证.数据质量问题包括两方面,下面将分别加以介绍.

一方面是知识缺失,比如 Wikipedia 页面中并不能涵盖所有实体,以及结构化知识库中缺失实体、关系及三元组.知识的缺失会影响到算法的精度,比如在相似实体搜索中,算法 QBEES,HeteSim 均未考虑知识缺失,算法的鲁棒性不好,精度没有考虑到知识缺失的 ARM 算法高.对于知识库中知识补全的方法主要包括利用对知识库进行张量分解的方法[72−77]以及表示学习方法[78−81],还有基于规则的补全方法[82].这些方法都是利用知识库本身来做补全,对于知识库中没有的实体、关系则不能补全.

另一方面就是知识错误问题.导致出现错误知识的原因很多,比如可能是语义的歧义、实体歧义、过时信息、单位错误等等.针对网页及知识库的错误知识发现目前有一些研究[83−87],这些方法主要集中在利用规则和基于统计两类方法.然而,知识库中有的错误是由于抽取规则造成的系统级错误,通过规则或者统计的方法是难以发现的.另外,目前已有的工作主要是针对某类错误提出的,比如文献[85]针对的是知识库三元组中宾语是数值型,根据统计发现数值的错误,在错误知识发现问题上还缺乏统一的模型.

References:

[1] Pound J, Mika P, Zaragoza H. Ad-Hoc object retrieval in the Web of data. In: Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 771−780. [doi: 10.1145/1772690.1772769]

[2] de Vries AP, Vercoustre AM, Thom JA, Craswell N, Lalmas M. Overview of the INEX 2007 entity ranking track. In: Proc. of the Focused Access to XML Documents, 6th Int'l Workshop of the Initiative for the Evaluation of XML Retrieval. Berlin: Springer-Verlag, 2007. 245−251. [doi: 10.1007/978-3-540-85902-4_22]

[3] Balog K, de Vries AP, Serdyukov P, Thomas P, Westerveld T. Overview of the TREC 2009 entity track. In: Proc. of the 18th Text REtrieval Conf. Gaithersburg: NIST, 2009.

[4] Unger C, Forascu C, Lopez V, Ngomo ACN, Cabrio E, Cimiano P, Walter S. Question answering over linked data (QALD-4). In: Proc. of the Working Notes for CLEF 2014 Conf. CEUR-WS.org, 2014. 1172−1180.

[5] Bizer C, Heath T, Idehen K, Berners-Lee T. Linked data on the Web (LDOW 2008). In: Proc. of the 17th Int'l Conf. on World Wide Web. New York: ACM Press, 2008. 1265−1266. [doi: 10.1145/1367497.1367760]

[6] Balog K, Azzopardi L, de Rijke M. Formal models for expert finding in enterprise corpora. In: Proc. of the 29th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2006. 43−50. [doi: 10.1145/1148170.1148181]

[7] Du F, Chen YG, Du XY. Survey of RDF query processing techniques. Ruan Jian Xue Bao/Journal of Software, 2013,24(6): 1222−1242 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4387.htm [doi: 10.3724/SP.J.1001.2013.04387]

[8] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives ZG. DBpedia: A nucleus for a Web of open data. In: Proc. of the Semantic Web, 6th Int'l Semantic Web Conf., 2nd Asian Semantic Web Conf. Berlin: Springer-Verlag, 2007. 722−735. [doi: 10.1007/978- 3-540-76298-0_52]

[9] Bollacker KD, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2008. 1247−1250. [doi: 10.1145/1376616.1376746]

[10] Hoffart J, Suchanek FM, Berberich K, Weikum G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. International Joint Conference on Artificial Intelligence, 2013,194:28−61. [doi: 10.1016/j.artint.2012.06.001]

[11] Dong XL, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, Strohmann T, Sun SH, Zhang W. Knowledge vault: A Web-scale approach to probabilistic knowledge fusion. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2014. 601−610. [doi: 10.1145/2623330.2623623]

[12] Zhang J, Tang J. Knowledge graph: The focus of next-generation search engine. Communications of the CCF, 2013,4(9):64−68 (in Chinese).

[13] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Trans. on Information Systems, 2002,20(4): 422−446. [doi: 10.1145/582415.582418]

[14] Yilmaz E, Kanoulas E, Aslam JA. A simple and efficient sampling method for estimating AP and NDCG. In: Proc. of the 31st Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2008. 603−610. [doi: 10.1145/1390334.1390437]

[15] Jiang LL. Studies on entity search and resolution [Ph.D. Thesis]. Lanzhou: Lanzhou University, 2012 (in Chinese with English abstract).

[16] Wang D, Niu JY. Entity retrieval method based on multi-perspective association model. Computer Engineering, 2013,1(39):71−75 (in Chinese with English abstract).

[17] Zhai CX. Statistical language models for information retrieval: A critical review. Foundations and Trends in Information Retrieval, 2008,2(3):137−213. [doi: 10.1561/1500000008]

[18] Kaptein R, Kamps J. Exploiting the category structure of Wikipedia for entity ranking. Artificial Intelligence, 2013,194:111−129. [doi: 10.1016/j.artint.2012.06.003]

[19] Chen YG, Gao LX, Shi SM, Du XY, Wen JR. Improving context and category matching for entity search. In: Proc. of the 28th Conf. on Artificial Intelligence. Palo Alto: AAAI, 2014. 16−22.

[20] Fang Y, Si L, Mathur AP. Discriminative models of integrating document evidence and document-candidate associations for expert search. In: Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2010. 683−690. [doi: 10.1145/1835449.1835563]

[21] Petkova D, Croft WB. Proximity-Based document representation for named entity retrieval. In: Proc. of the 2007 ACM CIKM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2007. 731−740. [doi: 10.1145/1321440.1321542]

[22] Lafferty JD, Zhai CX. Document language models, query models, and risk minimization for information retrieval. In: Proc. of the 24th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2001. 111−119. [doi: 10.1145/383952.383970]

[23] Balog K, Bron M, de Rijke M. Query modeling for entity search based on terms, categories, and examples. ACM Trans. on Information Systems, 2011,29(4):22. [doi: 10.1145/2037661.2037667]

[24] Liu TY. Learning to Rank for Information Retrieval. Berlin: Springer-Verlag, 2011. [doi: 10.1007/978-3-642-14267-3]

[25] Sorg P, Cimiano P. Finding the right expert—Discriminative models for expert retrieval. In: Proc. of the Int'l Conf. on Knowledge Discovery and Information Retrieval. Setúbal: SciTe Press, 2011. 190−199.

[26] Yang Z, Tang J, Wang B, Guo JY, Li JZ, Chen SC. Expert2Bólè: From expert finding to Bólè search. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2009. 1−4.

[27] Lin B, Rosa KD, Shah R, Agarwa N. LADS: Rapid development of a learning-to-rank based related entity finding system using open advancement. In: Proc. of the EOS, SIGIR 2011 Workshop, 2011.

[28] Hu GP, Liu JJ, Li H, Cao YB, Nie JY, Gao JF. A supervised learning approach to entity search. In: Proc. of the Information Retrieval Technology. Berlin: Springer-Verlag, 2006. 54−66. [doi: 10.1007/11880592_5]

[29] Macdonald C, Ounis I. Learning models for ranking aggregates. In: Proc. of the Advances in Information Retrieval—33rd European Conf. on IR Research. Berlin: Springer-Verlag, 2011. 517−529. [doi: 10.1007/978-3-642-20161-5_52]

[30] Lawrence P, Sergey B, Rajeev M, Terry W. The PageRank citation ranking: Bringing order to the Web. Technical Report, Stanford InfoLab, 1998.

[31] Kleinberg JM. Authoritative sources in a hyperlinked environment. In: Proc. of the 9th Annual Symp. on DIscrete Algorithm. New York: ACM Press, 1998. 668−677.

[32] Serdyukov P, Rode H, Hiemstra D. Modeling multi-step relevance propagation for expert finding. In: Proc. of the 2008 ACM CIKM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2008. 1133−1142. [doi: 10.1145/1458082.1458232]

[33] Serdyukov P, Rode H, Hiemstra D. Modeling expert finding as an absorbing random walk. In: Proc. of the 31st Annual Int'l Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2008. 797−798. [doi: 10.1145/1390334.1390509]

[34] Blanco R, Halpin H, Herzig DM, Mika P, Pound J, Thompson HS, Duc TT. Entity search evaluation over structured Web data. In: Proc. of the EOS. 2011. 65−71.

[35] Halpin H, Herzig DM, Mika P, Blanco R, Pound J, Thompson HS, Duc TT. Evaluating ad-hoc object retrieval. In: Proc. of the Int'l Workshop on Evaluation of Semantic Technologies. CEUR-WS.org, 2010.

[36] Neumayer R, Balog K, Norvag K. On the modeling of entities for ad-hoc entity search in the Web of data. In: Proc. of the Advances in Information Retrieval, 33rd European Conf. on IR Research. Berlin: Springer-Verlag, 2012. 133−145. [doi: 10.1007/978-3-642-28997-2_12]

[37] Balog K, Nørvåg K. On the use of semantic knowledge bases for temporally-aware entity retrieval. In: Proc. of the 5th Workshop on Exploiting Semantic Annotations in Information Retrieval. New York: ACM Press, 2012. 1−2. [doi: 10.1145/2390148.2390150]

[38] Zhiltsov N, Agichtein E. Improving entity search over linked data by modeling latent semantics. In: Proc. of the 2013 ACM CIKM Int'l Conf. on Information and Knowledge Management. New York: ACM Press, 2013. 1253−1256. [doi: 10.1145/2505515.2507868]

[39] Ding L, Pan R, Finin TW, Joshi A, Peng Y, Kolari P. Finding and ranking knowledge on the semantic Web. In: Proc. of the Semantic Web-ISWC 2005, 4th Int'l Semantic Web Conf. Berlin: Springer-Verlag, 2006. 156−170. [doi: 10.1007/11574620_14]

[40] Balmin A, Hristidis V, Papakonstantinou Y. ObjectRank: Authority-Based keyword search in databases. In: Proc. of the 30th Int'l Conf. on Very Large Data Bases. New York: ACM Press, 2004. 564−575. [doi: 10.1016/B978-012088469-8/50051-6]

[41] Nie ZQ, Zhang YZ, Wen JR, Ma WY. Object-Level ranking: Bringing order to Web objects. In: Proc. of the 14th Int'l Conf. on World Wide Web. New York: ACM Press, 2005. 567−574. [doi: 10.1145/1060745.1060828]

[42] Lim L, Wang HX, Wang M. Semantic queries by example. In: Proc. of the Joint 2013 EDBT/ICDT Conf. New York: ACM Press, 2013. 347−358. [doi: 10.1145/2452376.2452417]

[43] Bron M, Balog K, de Rijke M. Example based entity search in the Web of data. In: Proc. of the Advances in Information Retrieval, 35th European Conf. on IR Research. Berlin: Springer-Verlag, 2013. 392−403. [doi: 10.1007/978-3-642-36973-5_33]

[44] Takase S, Okazaki N, Inui K. Set expansion using sibling relations between semantic categories. In: Proc. of the 26th Pacific Asia Conf. on Language, Information and Computation. Stroudsburg: ACL, 2012. 525−534.

[45] Zhang L, Liu B. Entity set expansion in opinion documents. In: Proc. of the 22nd ACM Conf. on Hypertext and Hypermedia. New York, ACM Press, 2011. 281−290. [doi: 10.1145/1995966.1996002]

[46] Ghahramani Z, Heller KA. Bayesian sets. In: Proc. of the NIPS. 2005. 435−442.

[47] Letham B, Rudin C, Heller KA. Growing a list. Data Miningand Knowledge Discovery, 2013,27(3):372−395. [doi: 10.1007/s10618-013-0329-7]

[48] Gupta R, Sarawagi S. Answering table augmentation queries from unstructured lists on the Web. Proc. of the VLDB Endowment, 2009,2(1):289−300. [doi: 10.14778/1687627.1687661]

[49] Van Durme B, Pasca M. Finding cars, goddesses and enzymes: Parametrizable acquisition of labeled instances for open-domain information extraction. In: Proc. of the 23rd Conf. on Artificial Intelligence. Palo Alto: AAAI, 2008. 1243−1248.

[50] Sadamitsu K, Saito K, Imamura K, Kikui GI. Entity set expansion using topic information. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2011. 726−731.

[51] Li XL, Zhang L, Liu B, Ng SK. Distributional similarity vs. PU learning for entity set expansion. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2010. 359−364.

[52] Etzioni O, Cafarella MJ, Downey D, Kok S, Popescu AM, Shaked T, Soderland S, Weld DS, Yates A. Web-Scale information extraction in knowitall. In: Proc. of the 13th Int'l Conf. on World Wide Web. New York: ACM Press, 2004. 100−110.

[53] Lang J, Henderson J. Graph-Based seed set expansion for relation extraction using random walk hitting times. In: Proc. of the Human Language Technologies: Conf. of the North American Chapter of the Association of COmputational Linguistics. Stroudsburg: ACL, 2013. 772−776.

[54] Jindal P, Roth D. Learning from negative examples in set-expansion. In: Proc. of the 11th IEEE Int'l Conf. on Data Mining. Piscataway: IEEE, 2011. 1110−1115. [doi: 10.1109/ICDM.2011.86]

[55] Bing LD, Lam W, Wong TL. Wikipedia entity expansion and attribute extraction from the Web using semi-supervised learning. In: Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. New York, ACM Press, 2013. 567−576. [doi: 10.1145/2433396.2433468]

[56] Wang RC, Cohen WW. Automatic set instance extraction using the Web. In: Proc. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP. Stroudsburg: ACL, 2009. 441−449. [doi: 10.3115/1687878.1687941]

[57] Wang RC, Cohen WW. Language-Independent set expansion of named entities using the Web. In: Proc. of the 7th Int'l Conf. on Data Mining. Piscataway: IEEE, 2007. 342−350. [doi: 10.1109/ICDM.2007.104]

[58] Wang RC, Cohen WW. Iterative set expansion of named entities using the Web. In: Proc. of the 7th Int'l Conf. on Data Mining. Piscataway: IEEE, 2008. 1091−1096. [doi: 10.1109/ICDM.2008.145]

[59] Wang RC, Cohen WW. Character-Level analysis of semi-structured documents for set expansion. In: Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2009. 1503−1512. [doi: 10.3115/1699648.1699697]

[60] Wang RC, Schlaefer N, Cohen WW, Nyberg E. Automatic set expansion for list question answering. In: Proc. of the 2008 Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2008. 947−954. [doi: 10.3115/1613715.1613837]

[61] He YY, Xin D. SEISA: Set expansion by iterative similarity aggregation. In: Proc. of the 20th Int'l Conf. on World Wide Web. New York: ACM Press, 2011. 427−436. [doi: 10.1145/1963405.1963467]

[62] Sun YZ, Han JW, Yan XF, Yu PS, Wu TY. PathSim: Meta path-based top-*K* similarity search in heterogeneous information networks. PVLDB, 2011,4(11):992−1003.

[63] Passant A. DBREC—Music recommendations using DBpedia. In: Proc. of the 9th Int'l Semantic Web Conf. (ISWC 2010). Berlin: Springer-Verlag, 2010. 209−224. [doi: 10.1007/978-3-642-17749-1_14]

[64] Metzger S, Schenkel R, Sydow M. Aspect-Based similar entity search in semantic knowledge graphs with diversity-awareness and relaxation. In: Proc. of the 2014 IEEE Int'l Joint Conf. on Web Intelligence and Intelligent Agent Technologies. Piscataway: IEEE, 2014. 60−69. [doi: 10.1109/WI-IAT.2014.17]

[65] Abedjan Z, Naumann F. Improving RDF data through association rule mining. Datenbank-Spektrum, 2013,13(2):111−120. [doi: 10.1007/s13222-013-0126-x]

[66] Cai QQ, Yates A. Large-Scale semantic parsing via schema matching and lexicon extension. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2013. 423−433.

[67] Yao XC, Van Durme B. Information extraction over structured data: Question answering with freebase. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL 2014. 956−966. [doi: 10.3115/v1/P14-1090]

[68] Lee JY, Min JK, Oh A, Chung CW. Effective ranking and search techniques for Web resources considering semantic relationships. Information Processing and Management, 2014,50(1):132−155. [doi: 10.1016/j.ipm.2013.08.007]

[69] Lakkaraju H, Bach SH, Leskovec J. Interpretable decision sets: A joint framework for description and prediction. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2016. 1675−1648. [doi: 10.1145/2939672.2939874]

[70] Huang JZ, Zhao SQ, Ding SQ, Wu HY, Sun MM, Wang HF. Generating recommendation evidence using translation model. In: Proc. of the 26th Int'l Joint Conf. on Artificial Intelligence (IJCAI). Palo ALto: AAAI, 2016. 2810−2816.

[71] Yin H, Hu Z, Zhou X, Wang H, Zheng K, Nguyen QVH, Sadiq S. Discovering interpretable geo-social communities for user behavior prediction. In: Proc. of the 32nd Int'l Conf. on Data Engineering. Piscataway: IEEE, 2016. 942−953. [doi: 10.1109/ICDE.2016.7498303]

[72] Kolda T, Bader B. The TOPHITS model for higher-order Web link analysis. In: Proc. of the Workshop on Link Analysis Counterterrorism & Security. 2006.

[73] Kolda TG, Bader BW, Kenny JP. Higher-Order Web link analysis using multilinear algebra. In: Proc. of the 5th IEEE Int'l Conf. on Data Mining. Piscataway: IEEE, 2005. 242−249. [doi: 10.1109/ICDM.2005.77]

[74] Franz T, Schultz A, Sizov S, Staab S. TripleRank: Ranking semantic Web data by tensor decomposition. In: Proc. of the 8th Int'l Semantic Web Conf. (ISWC 2009). Berlin: Springer-Verlag, 2009. 213−228. [doi: 10.1007/978-3-642-04930-9_14]

[75] Drumond L, Rendle S, Schmidt-Thieme L. Predicting RDF triples in incomplete knowledge bases with tensor factorization. In: Proc. of the ACM Symp. on Applied Computing. New York: ACM Press, 2012. 326−331. [doi: 10.1145/2245276.2245341]

[76] Nickel M, Tresp V, Kriegel HP. A three-way model for collective learning on multi-relational data. In: Proc. of the 28th Int'l Conf. on Machine Learning. WI: Omnipress, 2011. 809−816.

[77] Krompass D, Nickel M, Tresp V. Large-Scale factorization of type-constrained multi-relational data. In: Proc. of the Int'l Conf. on Data Science and Advanced Analytics. Piscataway: IEEE, 2014. 18−24. [doi: 10.13140/2.1.4511.7441]

[78] Bordes A, Usunier N, García-Durán A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Proc. of the Advances in Neural Information Processing Systems 26, 27th Annual Conf. on Neural Information Processing Systems 2013. Berlin: Springer-Verlag, 2013. 2787−2795.

[79] Wang Z, Zhang JW, Feng JL, Chen Z. Knowledge graph embedding by translating on hyperplanes. In: Proc. of the 29th Conf. on Artificial Intelligence. Palo Alto: AAAI, 2014. 1112−1119.

[80] Lin YK, Liu ZY, Sun MS, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. In: Proc. of the 29th Conf. on Artificial Intelligence. Palo Alto: AAAI, 2015. 2181−2187.

[81] Socher R, Chen DQ, Manning CD, Ng AY. Reasoning with neural tensor networks for knowledge base completion. In: Proc. of the Advances in Neural Information Processing Systems 26, 27th Annual Conf. on Neural Information Processing Systems 2013. Berlin: Springer-Verlag, 2013. 926−934.

[82] Chen Y, Goldberg SL, Wang DZ, Johri SS. Ontological pathfinding. In: Proc. of the 2016 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2016. 835−846. [doi: 10.1145/2882903.2882954]

[83] Meusel R, Paulheim H. Heuristics for fixing common errors in deployed schema.org microdata. In: Proc. of the Semantic Web: Latest Advances and New Domains. Berlin: Springer-Verlag, 2015. 152−168. [doi: 10.1007/978-3-319-18818-8_10]

[84] Paulheim H, Bizer C. Improving the quality of linked data using statistical distributions. Int'l Journal of Semantic Web Information Systems, 2014,10(2):63−86. [doi: 10.4018/ijswis.2014040104]

[85] Wienand D, Paulheim H. Detecting incorrect numerical data in DBpedia. In: Proc. of the Semantic Web: Trends and Challenges, 11th Int'l Conf. Berlin: Springer-Verlag, 2014. 504−518. [doi: 10.1007/978-3-319-07443-6_34]

[86] Li X, Dong XL, Lyons K, Meng WY, Srivastava D. Truth finding on the deep Web: Is the problem solved? Proc. of the VLDB Endowment, 2012,6(2):87−108.

[87] Li X, Dong XL, Lyons KB, Meng WY, Srivastava D. Scaling up copy detection. In: Proc. of the 31st Int'l Conf. on Data Engineering. Piscataway: IEEE, 2015. 89−100. [doi: 10.1109/ICDE.2015.7113275]

附中文参考文献:

[7] 杜方,陈跃国,杜小勇.RDF 数据查询处理技术综述软件学报,2013,24(6):1222−1242. http://www.jos.org.cn/1000-9825/4387.htm [doi: 10.3724/SP.J.1001.2013.04387]

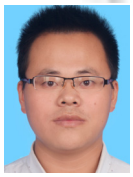
[12] 张静,唐杰.下一代搜索引擎的焦点:知识图谱.中国计算机学会通讯,2013,4(9):64−68.

[15] 姜丽丽.实体搜索与实体解析方法研究[博士学位论文].兰州:兰州大学,2012.

[16] 王东,牛军钰.基于多角度关联模型的实体检索方法.计算机工程,2013,1(39):71−75.

张香玲(1983－),女,山东滨州人,博士生,CCF 学生会员,主要研究领域为实体检索,知识补全.

陈跃国(1978－),男,博士,副教授,博士生导师,CCF 高级会员,主要研究领域为交互式大数据分析,实体搜索.



马登豪(1989－),男,博士生,主要研究领域为信息检索,实体搜索.

陈峻(1991－),男,博士生,CCF 学生会员,主要研究领域为探索式搜索,信息检索,大数据管理.

杜小勇(1963－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据库系统,智能信息检索.