基于多语本体的语义查询扩展研究*

司莉[1]，潘秋玉[2]
（1. 武汉大学信息资源研究中心，武汉 430072；2. 武汉大学信息管理学院，武汉 430072）

摘要：查询扩展是改善信息检索结果的有效方法。针对用户获取多语言信息的需求以及当前跨语言信息检索存在的翻译歧异性问题，提出一种基于多语本体的语义查询扩展方法，介绍其基本原理、查询扩展模型及实现过程，使跨语言信息检索从字符匹配变成语义层面的匹配，实现跨语言信息检索中的查询扩展，以提高多语言信息检索的查全率和查准率。

关键词：查询扩展；多语本体；跨语言信息检索
中图分类号：TP391
DOI：10.3772/j.issn.1673-2286.2016.2.006

1 引言

查询扩展作为提高信息检索性能的关键技术，自20世纪60年代提出以来就逐渐受到关注。尤其是近年来，在计算机技术、云计算、物联网、用户创造内容等多重因素的推动下，Internet已成为一个海量且仍在迅猛增长的信息库，与此同时，网络信息语种的多样化和网络用户分布的国际化日益显著，实现多语言信息组织与检索，使用户方便获取多语言信息，成为信息检索系统发展的趋势之一。然而，不同语言概念之间的准确对应始终是制约多语言信息检索的瓶颈。由于本体表达概念语义和推理的能力较强，可消除自然语言理解中的歧义，并能根据相关概念进行推理，在多语言信息检索中实现基于本体的语义查询扩展，将有效提高多语言信息检索的查全率和查准率，从而促进全球知识交流与共享。

2 现有的查询扩展技术

传统信息检索系统利用简单的词匹配法则，即计算文档特征值与检索词之间的相似度，往往只能检索到包含查询词的那些资源。而用户输入的89.9%的检索查询只包含一个词，平均查询词为1.73个[1]，这样，与用户查询请求相关但未包括检索词的那部分资源便无法被命中。可见，实现用以提高查全率的查询扩展是极为必要的。查询扩展的基本思想是对用户输入的初始查询词进行修正和扩充，构建更明确清晰的查询表达式，以改善信息检索的查全率和查准率。扩充的查询词有两大类：一是查询词的同义或近义词，二是加入全新的词汇。

目前，查询扩展的常用方法有3种：①基于用户相关反馈的查询扩展；②基于全局分析的查询扩展；③基于局部分析的查询扩展。其中，第一种要求用户对查询结果进行相关性判断，系统对用户判断后的相关文档进行计算，选取一些词扩展查询式进行二次检索，如此反复直至用户满意，该方法可以很好地满足用户需求，但对用户要求较高、负担较重；第二种是系统自动对全部文档中的词或词组进行相关分析，将与查询词关联度较高的词作为初始查询词的扩展词来生成新的查询式，其缺点是当文档数量较多时，计算量会比较大；第三种不需要用户参与，系统自动将查询结果中的前K篇文档作为相关文档，计算后选取扩展词进行重新检索，但容易发生“查询漂移”现象，即扩展后的查询主题偏离了用户原来的查询意图。

3 基于多语本体的查询扩展方法

3.1 多语本体的特征

本体是对概念及概念之间关系规范化、形式化、可

* 本研究得到教育部人文社会科学重点研究基地重大项目“基于内容的多语言信息组织与检索研究”（编号：14JJD870001）资助。

共享、明确化的描述，是一种表达、共享、重用知识的方法[2]。多语本体是本体在不同语种中的具体表现形式[3]。多语本体不同于多语种词典，因为它不仅包含大量规范的多语种概念，还具有丰富的概念关系和强大的推理能力。除了具备本体的一般特征，多语本体还有一个重要特征，即多语言同义词规范。多语本体库中的概念虽在各语言中的表达方式不同，但它们的内涵是相互对应且一致的。词汇只是概念的一种表达方式，概念是独立于语言的，所以一个概念的内涵甚至可以不用语言表示，而使用数字或者符号等来代替。

目前已建立了多语本体WordNet以及以WordNet为标准建立的多语本体系列，如欧洲的EuroWordNet、中国的HowNet、俄国的RussianWordNet等。其中，EuroWordNet包含英语、荷兰语、意大利语、西班牙语、德语、法语、捷克语、爱沙尼亚语等八个部分，它们之间通过中间语言索引将一种语言中的概念与其他语言中相似的概念联系起来[4]；HowNet是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，描述了上下位关系、同义关系、反义关系、部件-整体关系等16种词间关系[5]。多语本体的构建为多语言语义检索提供支持，是多语言信息检索研究中重要的语言资源和工具。

3.2 基于多语本体的查询扩展

3.2.1 基本原理

传统的跨语言信息检索最常用的方法是提问式翻译，即将用户输入的提问式翻译为系统支持的其他每种语言，然后进行单语言检索[6]。这种方法的缺点是提问式往往没有语境支持，这种简单的关键词翻译难以避免翻译过程中的歧义性问题。由于多语本体具有丰富的概念关系和强大的推理能力，使得基于多语本体的查询扩展能够将提问式与文档的对照和匹配提升到语义层面，从而有效地完成消歧工作。

在多语本体中，不同语种的概念术语通过映射进行了关联，当用户输入一种语言的查询语句时，系统在源语言本体库中检索相应的查询结果，并自动映射到其他语种，搜索与目标语言概念相同或相近的结果反馈给用户。多语本体在多语言信息检索的作用主要体现在两个方面：一是在转换查询语言时，对提问式进行分词和概念提取，并与多语本体库中的内涵进行对比，根据不同对应情况作不同处理；二是在多语言信息检索时，对检索对象进行语义层面的处理，计算潜在文档与查询提问式之间的语义相关性，并按从高到低的顺序排列，将查询结果返还给用户。

3.2.2 基于多语本体的查询扩展模型

定义用户的初始查询为Q，对Q进行分词、提取概念等预处理后可表示为$Q=\{q_1,q_2,q_3,\cdots,q_n\}$，然后判断$Q=\{q_1,q_2,q_3,\cdots,q_n\}$的查询模式类型，按照不同的算法借助多语本体库进行语义扩展，得到候选查询扩展词集$Q_e=\{q_{n+1},q_{n+2},q_{n+3},\cdots,q_{n+n}\}$，计算初始查询词$q_i$与$Q_e$中的每个候选扩展词$q_{n+i}$之间的相似度$sim(q_i, q_{n+i})$。为了避免查询扩展词过多而影响查询扩展的精度和检索结果的查准率，引入阈值λ（通过实验得到）来对扩展词进行一定的控制。比较每个$sim(q_i, q_{n+i})$值与给定阈值λ的大小，并保留Q_e中$sim(q_i, q_{n+i})>\lambda$的扩展词$q_{n+i}$，作为最终的查询扩展词集。根据以上描述，基于多语本体的查询扩展模型如图1所示。

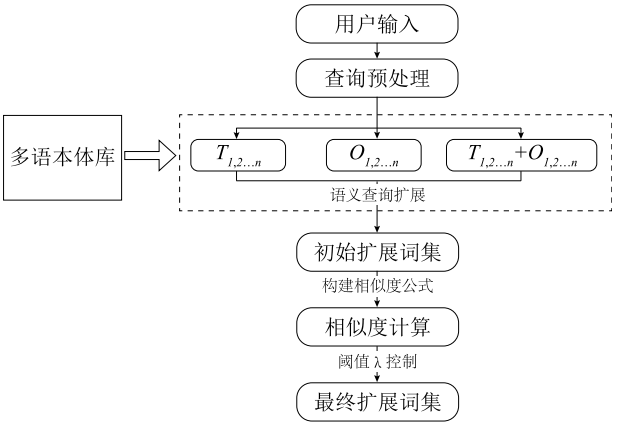

图1 基于多语本体的查询扩展模型

（1）查询预处理

即接收用户输入的查询词，并进行分词、切词、句法语义关联分析、提取概念、去除停用词及多余符号等预处理，得到有检索意义的关键词集合$Q=\{q_1,q_2,q_3,\cdots,q_n\}$。在预处理查询之前，应当在一定程度上了解用户的查询行为。如在查询词方面，用户输入查询词时虽各有偏好，但大致可归纳为三类：①单个关键词查询；②多个关键词查询；③自然语言查询，可分

别以“杜鹃花”“湖滨杜鹃花”“武汉大学湖滨有杜鹃花吗？”为例。了解用户的查询行为，有助于对用户输入的初始查询词进行有效处理。

（2）语义查询扩展

对用户查询预处理后，判断$Q=\{q_1,q_2,q_3,\cdots,q_n\}$的查询模式类型。用户的查询模式主要分为3种类型：一是$T_{1,2,\ldots,n}$模式，即用户输入的关键词$Q=\{q_1,q_2,q_3,\cdots,q_n\}$不是多语本体中的概念或者实例；二是$O_{1,2,\ldots,n}$模式，即用户输入的关键词$Q=\{q_1,q_2,q_3,\cdots,q_n\}$是多语本体中的概念或者实例；三是$T_{1,2,\ldots,n}+O_{1,2,\ldots,n}$模式，也称混合模式，即用户输入的关键词$Q=\{q_1,q_2,q_3,\cdots,q_n\}$既有多语本体中的概念或者实例，也包含不在本体库中的词汇[7]。根据用户不同的查询模式，借助多语本体运用不同方法完成语义扩展。

①$T_{1,2,\ldots,n}$模式的语义查询扩展

查询关键词不在多语本体中，使用词典翻译关键词t_i，采用基于关键词匹配的方法检索，获取每一个t_i的相关文档，并统计这些文档中出现的t_i和本体概念及其各自出现的频次，选择前n个本体概念作为扩展概念，并完成对普通关键词的扩展。其基本思想是每个关键词常常会出现在某个相应的语境中，在该语境中同时出现的词往往与查询关键词有着密切的关联，同理，在该语境中出现的本体概念也与查询关键词有某种联系[8]。通过这样的方式把用户输入的普通关键词语义化，这些本体概念作为关键词的扩展词也有着相当的语义价值。

②$O_{1,2,\ldots,n}$模式的语义查询扩展

直接将查询词与多语本体库中概念的内涵进行映射，找出合适的本体概念以及相关的术语、关系、实例等。由于在多语本体中，不同语种的概念术语通过映射进行了关联，当用户输入一种语言的查询语句时，系统在源语言本体库中检索对应结果，系统可以自动映射到其他语种，搜索与目标语言概念相同或相近的结果反馈给用户。例如，若以中文、英文和日文建立珞珈山植物多语本体库，用户输入中文关键词“映山红”，系统首先调用中文库，与本体中的术语进行匹配，把“杜鹃”“山石榴”“唐杜娟”等同义词汇选出来，再把与这些词汇相关的上级概念、同类概念、地理分布等关系找出来；利用多语本体的映射关系找出英文、日文中对应的术语及其相关概念，如“rhododendron”“ツツジ”等，系统以"映山红"及其中文、英文、日文三种语言的扩展词进行检索，从而实现语义查询扩展和多语言信息检索。

③$T_{1,2,\ldots,n}+O_{1,2,\ldots,n}$模式的语义查询扩展

此模式是上述两种模式混合的情况。在用户的查询中既有多语本体中的概念，也有多语本体不能直接处理的普通关键词。这种模式有两种情况，第一种是$T_{1,2,\ldots,n}$中的信息与$O_{1,2,\ldots,n}$中的属性的取值相关，第二种是$T_{1,2,\ldots,n}$中的信息与$O_{1,2,\ldots,n}$中的属性的取值并不相关[8]。仍以珞珈山植物多语本体库为例，如在“湖滨有杜鹃花吗？”查询中，“杜鹃花”是多语本体库中的概念，“湖滨”是杜鹃花地理分布范围的值，可在多语本体库中找出这一关系，返还给用户相关文档；而在“rhododendron DuFu”查询中，“rhododendron”是多语本体库中的概念，“DuFu”则不在本体库中，运用多语本体库扩展出“rhododendron”的相关词“杜鹃”“山石榴”“唐杜娟”“ツツジ”等（$O_{1,2,\ldots,n}$模式），使用词典找到“DuFu”的对应翻译词“杜甫”（$T_{1,2,\ldots,n}$模式），再使用“杜甫”与“rhododendron”及其扩展词匹配检索，返回用户需求的信息。

（3）语义相似度计算

语义相似度是指两个词语在语义层次上的相似程度，即它们在上下文语境中能够在不改变句法的前提下相互交换的程度[9]，其取值在[0,1]之间，两个完全相同的词语语义相似度为1，如“映山红”和“杜鹃”；两个不能互相代替的词语语义相似度为0，如“映山红”和“杜甫”等。

当要准确计算出两个概念间的相似度时，首先必须清楚影响语义相似度的因素，主要有：①语义距离$Dis(X,Y)$，即两个概念X、Y在层次网中的距离，一般用两个概念各自对应的节点在层次树中的最短路径来衡量。语义距离越小，两个概念间的语义相似度越大，$Sim(X,Y)$值越接近于1，相反，语义距离越大，概念间的语义相似度越小，$Sim(X,Y)$值越接近于0，两者呈反比关系；②概念节点的深度$Depth(X)$，即概念X在本体中与根节点的最短长度，$Depth(X)=Lenth(root,X)$，这里有两种情境，一是在本体中处于同一层次的概念间的相似度大于不同层次的概念间的相似度，二是当语义距离相同时，在本体层次树中距离根节点远的概念节点间的相似度大于离根节点近的概念节点间的相似度；③概念节点的宽度$Width(X)$，即概念X在本体中同一层次概念子节点的数量，子节点数越多，说明细化程度越大，分类越具体，概念间的语义相似度也越大，反之则越小；④语义重合度$Match(X,Y)$，即本体内部两概念结点X、Y之间包含相同的上位概念在总节点中所占的比例，其基本思想是两个概念拥有共同父节点的数量越

多，表明两者关联度越高，相似度越大。

（4）阈值控制

即引入阈值λ对扩展词的数量进行一定的控制，以保证查询扩展的精度。阈值λ的值需要通过实验获取。利用语义相似度计算公式计算初始查询词q_i与每个候选扩展词q_{n+i}之间的相似度$sim(q_i, q_{n+i})$，删除相似程度低于阈值λ的候选扩展词，同时保留概念间语义相似度值大于阈值λ的词汇，形成最终的查询扩展词集。这样不仅对用户输入的查询词进行了语义层面的操作，而且通过限制搜索范围避免了查准率降低的问题，从而使查询扩展更符合用户需求，保证检索结果的全面性和精确性。

3.2.3 基于多语本体的查询扩展实现过程

综上，具体的基于多语本体的查询扩展实现过程如下：

（1）利用多语词典等相关资源和本体构建工具建立一个多语言领域本体库；

（2）用户输入查询式，系统对查询式进行分词、去除停用词、提取概念等处理，把用户查询式表示为$Q=\{q_1,q_2,q_3,\cdots, q_n\}$；

（3）根据$Q=\{q_1,q_2,q_3,\cdots, q_n\}$所属的查询模式类型及其各自的查询扩展算法，借助多语本体库进行语义扩展，将查询词与源语言本体库中概念的内涵进行映射，找出合适的本体概念以及相关的术语，并自动映射到其他语种，查找其他语言中相对应的概念，得到包含各语种的查询扩展词集$Q_e=\{q_{n+1},q_{n+2},q_{n+3},\cdots, q_{n+n}\}$；

（4）利用语义相似度公式计算出初始查询词q_i与每个候选扩展词q_{n+i}之间的相似度$sim(q_i, q_{n+i})$，并与阈值λ比较，把$sim(q_i, q_{n+i})>\lambda$的词汇加入到扩展词集中；

（5）将最终查询式$Q\cup Q_e$提交给搜索引擎实施检索。

4 结语

笔者在传统跨语言信息检索的基础上提出一种基于多语本体的查询扩展方法，描述了多语本体在跨语言信息检索中的应用原理，建立并详细介绍了基于多语本体的查询扩展模型，使跨语言信息检索由关键词匹配进化为语义匹配，能够在一定程度上改善信息检索性能，实现多语言信息检索的语义扩展，有效提高获取全球知识的效率。将该方法运用于跨语言信息检索的前提是要建立一个优秀的多语本体库，并设计合适的算法，本文尚未使用实际的系统进行实验以验证该方法的有效性，有待我们在下一步的研究工作中进行实现。

参考文献

[1] 胡保祥.基于查询日志的查询扩展研究[D].北京:北京邮电大学,2013.

[2] 司莉.信息组织原理与方法[M].武汉:武汉大学出版社,2011:269.

[3] 吴丹,王惠临.本体在跨语言信息检索中的应用机制研究[J].图书情报工作,2006,50(9):10-13.

[4] Vossen P J. EuroWordNet: Building a multilingual database with wordnets for several European languages. [EB/OL]. [2015-11-20]. http://www.illc.uva.nl/EuroWordNet/.

[5] 董振东.《知网》中文版[EB/OL]. [2015-11-20]. http://www.keenage.com/html/c_index.html.

[6] 吴丹.本体驱动的跨语言信息检索研究[J].现代图书情报技术,2006(5):22-26,85.

[7] 王进,陈恩红,张振亚,等.基于本体的跨语言信息检索模型[J].中文信息学报,2004(3):1-8,60.

[8] 高敏.基于本体的语义查询扩展研究[D].济南:山东科技大学,2008.

[9] 谭睿哲.基于本体和用户日志的查询扩展研究[D].长沙:湖南大学,2013.

作者简介

司莉，女，1965年生，武汉大学信息资源研究中心研究员，武汉大学信息管理学院教授、博导、图书馆学系主任，研究方向：信息组织、知识组织、图书馆营销与服务等，E-mail：lsiwhu@163.com。

潘秋玉，女，1991年生，武汉大学信息管理学院硕士研究生，研究方向：信息组织。

Semantic Query Expansion Based on Multilingual Ontology

SI Li[1], PAN QiuYu[2]
(1. The Center for the Study of Information Resources, Wuhan University, Wuhan 430072, China;
2. School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: Query expansion is an effective method to enhance information retrieval performance. Aiming at the requirements of acquiring multilingual information and solving the problems of semantic disambiguation of cross language information retrieval (CLIR), the article proposed a new semantic query expansion method based on multilingual ontology, and introduced its fundamentals, model and realization process, to turn character-matching into semantic matching for CLIR, implementing query expansion in CLIR, which may optimize system’ s recall and precision.

Keywords: Query Expansion; Multilingual Ontology; Cross Language Information Retrieval (CLIR)

（收稿日期：2016-01-15）

2016年全国知识组织与知识链接学术交流会征文通知

为了探讨网络环境下知识组织与知识链接的新理念、新思路、新方法，中国科学技术信息研究所、国家科技图书文献中心、中国科学技术情报学会拟于2016年下半年召开第七届“全国知识组织与知识链接学术交流会”，由《数字图书馆论坛》承办。特向国内图书情报界及相关领域的专家学者征文。具体会议时间、地点另行通知。

一、会议主题

知识组织

- 知识组织研究进展
- 大数据中的数据清洗、组织和分析
- 事实型数据识别与分析
- 叙词表、本体等知识组织体系的构建和应用研究

知识评价

- 科学计量与评价
- 引文分析、主题分析
- 专利分析与利用
- Web科技资源评价

知识链接

- 科研实体关系揭示
- 网络资源链接及其关联分析
- 面向项目研发产出的关联分析
- 数据关联挖掘和揭示

知识服务

- 知识服务、知识管理研究进展
- 数字科研环境与开放共享服务
- 用户分析与个性化用户服务
- 知识图谱及可视化分析

二、联系方式

会议邮箱：KOLink@istic.ac.cn
会议网址：http://168.160.16.186/conference