面向跨领域海量信息资源的元搜索引擎研究

朝乐门　张　勇　邢春晓

摘　要　跨领域海量信息资源搜索是从事交叉学科和跨学科研究的重要前提。元搜索引擎不仅可以避免独立搜索引擎覆盖面较低的局限性,而且还可以充分发挥不同独立搜索引擎之间的互补性。基于元搜索引擎的跨领域海量信息资源搜索需要解决四个关键问题,即协助用户准确表达搜索意图、消除检索词的“一词多义”和“多词一义”现象、设计自动调整检索范围的机制以及发挥用户价值。面向跨领域海量信息资源的元搜索引擎采用多领域本体、语义 Web 和 Web2.0 技术,较好地解决了元搜索引擎的上述四个关键问题。相对于传统元搜索引擎,面向跨领域海量信息资源的元搜索引擎在基本思路、框架设计、流程设计、关键技术四个方面发生了重要变化。图 4。表 1。参考文献 36。

关键词　元搜索引擎　海量信息资源　多领域本体　语义 Web　Web2.0

分类号　TP182

ABSTRACT　It is essential for cross discipline researchers to search massive cross-domain information resources. Meta-search engines are widely used for that purpose. However, today' s meta-search engines are facing four challenges: to help users express accurately their search intentions, to overcome vocabularies mismatching of polysemy and synonym, to adjust automatically the search scope, and to maximize user-added values. Based on technologies of multiple domain ontologies, the Semantic Web and Web2.0, a new search engine for massive cross-domain information resources is proposed and may provide possible solutions to the above four problems. Compared to the traditional meta-search engines, the new one makes four important changes in terms of basic principles, framework design, main procedures and key technologies. 4 figs. 1 tab. 36 refs.

KEY WORDS　Meta-search engine.　Information resource.　Multiple domain ontology.　Semantic web.　Web2.0.

CLASS NUMBER　TP182

1　引言

跨领域海量信息资源搜索是从事交叉学科研究、推动新兴学科领域发展的重要前提。目前,独立搜索引擎已成为获取跨领域海量信息资源的主要手段。独立搜索引擎具有两个基本特征:第一,单个独立搜索引擎的覆盖面较低。文献[1]的调查研究证明独立搜索引擎仅能返回所有 WWW 资源的 15%;第二,两个不同搜索引擎对同一个检索提问的重复率较低。文献[2]对 MSN、Google、Yahoo、Ask Jeeves 等四大搜索引擎对同一个搜索提问所返回结果进行比较研究发现,其中四个、三个、两个搜索引擎中均返回的搜索结果的重复率分别为 1.1%、2.6% 和 11.4%。根据独立搜索引擎的上述两个基本特征,可以得出如下两条结论:第一,由于独立搜索引擎的覆盖面有限,任何一个独立搜索引擎都无法胜任跨领域海量信息资源的搜索任务;第二,由于独立搜索引擎之间的重复率低,多个搜索引擎的搜索结果具有互补性,通过多个独立搜索引擎的集成应用可以较好地实现跨领域海量信息资源的搜索任务。因此,本文主要探讨基于元搜索引擎的海量信息资源搜索及其改进方案。论文的主要内容安排如下:第二部分主要梳理了现有研究基础以及目前面向跨领域海量信息资源的元搜索引擎研究中需要解决的四个主要挑战;第三部分以解决上一部分

所提出的四个挑战为主要目的,设计了面向跨领域海量信息资源的元搜索引擎研究的基本思想、框架设计、流程设计和关键技术;第四部分采用 OWL 测试领域本体集和 Jena 接口开发出了原型系统,验证了本文研究的可行性;最后对论文研究进行简要总结,并描述了下一步研究工作。

2 相关研究

2.1 研究现状

从以上分析看出,元搜索引擎可以弥补独立搜索引擎的两个局限性,为跨领域海量信息的搜索提供了重要解决方案。元搜索引擎是指一种通过转发用户搜索提问至多个独立搜索引擎、Web 目录或尚未被传统独立搜索引擎直接索引的隐藏网络(Deep Web),并对所返回结果进行重复过滤、合并、排序等以完成用户提交的一次信息检索任务的搜索工具[3]。与传统搜索引擎不同的是,元搜索引擎一般不需要通过自己的网络机器人爬取网络信息资源和在本地存储与维护网络资源的索引库。因此,相对于独立搜索引擎,元搜索引擎具有覆盖面广、维护方便等优势。元搜索引擎一般由三部分组成,即检索请求提交代理、检索接口代理、检索结果显示代理[4](见图 1)。元搜索引擎的工作过程分为如下步骤:"检索请求代理"负责接收用户的原始查询,并把原始查询分别转换为各个成员搜索引擎能够接受的形式;"检索接口代理"负责向成员搜索引擎发送查询请求;"结果显示代理"负责收集各个搜索引擎的原始查询结果,并对其结果进行合并、去重和排序,把最终查询结果递交给用户[5]。

元搜索引擎在理论研究和开发应用领域取得了一定的进展。学术界在元搜索引擎的基本原理、组成部分和工作过程等方面已达成共识,成员选择、相关性排序、个性化、效率提高逐渐成为该领域学术研究的热点问题。其中,成员选择机制可分为两种,即系统选择机制和用户选择机制,用户选择机制可以根据用户具体需求选择不同的成员搜索引擎,能够提高信息资源搜索的准确率[6];相关性排序技术可分为收集结果重新排序(直接合并、根据相应速度排序、摘要排序、成员排序)、利用搜索引擎排序信息排序(轮询法、星星排序、Borda 排序、位置排序、概念可信度排序、贝叶斯概率模型排序)和相关分值融合(Comb 排序、SDM 法、MEM 法、CORINET 法)三大类[7];元搜索引擎的个性化需要在元搜索引擎的三个基本组成部分的基础上,再增加个性化代理和个人信息库,用来计算用户个人信息或其搜索历史记录与检索结果之间的相关性,并将相关性较高的检索结果显示给用户。文献[8]提出一种基于用户记录(User Profile)的元搜索引擎方案,探讨了如何根据用户记录信息推理用户信息需求的问题;文献[9]探讨了如何通过用户提问、检索结果和点击记录来提高元搜索引擎的查准率问题。在应用开发领域,诸多元搜索引擎已被投入使用,元搜索引擎之间的比较分析成为热点话题。文献[10]采用计算元搜索引擎与所对应的独立搜索引擎之间的距离(closeness)方法,对 Clusty、Dogpile、Excite、Mamma、MetaCrawler、Search. com、WebCrawler 和 Webfetch 等八个典型元搜索引擎的搜索效果进行了评估;文献[11]从信息检索功能(布尔检索、词组短语检索、截词检索、限定检索、位置检索、概念检索、多语种检索和自然语

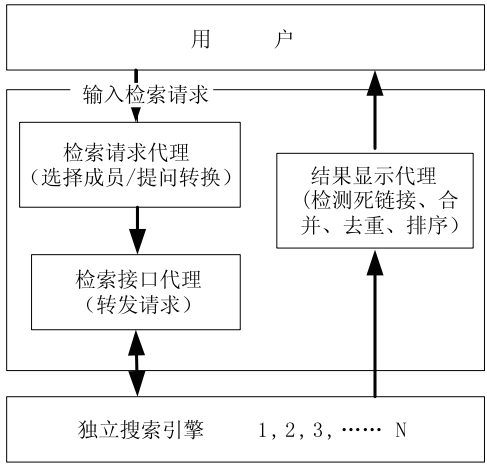

图 1 元搜索引擎原理

言检索)的角度比较分析了万纬、搜星、一起搜、Bbmao、Jux2、Dogpile、Mamma、Clusty 和 Vivisiomo 等元搜索引擎,发现元搜索引擎的限定检索、位置检索、概念检索、多语种检索、自然语言检索有待进一步改进。

为了提高元搜索引擎的查准率,领域元搜索引擎(Domain Meta Search Engine)成为元搜索引擎的重要研究趋势之一。从目前的研究现状看,领域元搜索引擎主要通过将领域本体技术引入到元搜索引擎体系结构之中,利用领域本体中的概念及其概念之间的关系,实现用户查询关键词的同义词扩展、英文扩展以及改进摘要排序算法[12]、用户意图与领域本体的匹配[13]、查询结果处理[14]、上下文敏感检索[15]、文本内容的理解[16],进而提高元搜索引擎的查全率和查准率。

2.2　研究挑战

虽然元搜索引擎,尤其是领域元搜索引擎对跨领域海量信息资源搜索提供了较好的解决方案,但是仍未解决如下几个重要问题:

(1)用户准确表达自己的搜索意图的困难。用户搜索意图的准确表示是提高元搜索引擎查全率和查准率的重要前提条件。在基于元搜索引擎的跨领域信息资源搜索过程中,用户需要选用适当的关键词或提问表达式(Q)表示自己的搜索意图(R)。受搜索用户的信息搜索意图表达能力所限,用户选择的关键词可能是用户搜索意图(R)的上位概念、下位概念或相关概念,而并不是概念本身,导致搜索意图(R)向提问表达式(Q)转换过程中信息扭曲现象的出现,影响元搜索引擎的后续处理效果。

目前,解决用户准确表示其搜索意图的方法是采用用户搜索意图的智能化理解系统。对用户搜索意图的智能化理解需要检索系统(软件)从概念层次理解用户提问,现有两种做法:一是基于同义词表、蕴含词表等入口词表辅助进行领域和概念的扩充;二是利用自然语言处理技术,对提问加以分析[17]。但是,二者均有其局限性,具体表现在两个方面:一是由于同义词表、蕴涵表中的概念及其概念之间的关系的表示能力有限,无法以形式化的方式直接表示类概念之间的继承、等同、交叉关系,属性之间的继承、等同和互逆关系,属性与类概念之间的对应关系,类与实例之间的对应关系,概念之间的传递、对称、函数、反函数关系,概念之间的代数运算关系,概念之间的不同版本关系等;二是由于人工智能领域在知识处理领域的突破性进展比较缓慢,使得用自然语言实现计算机对用户检索提问的理解非常困难。因此,如何通过规范化表示不同领域的知识及知识之间的复杂关系,不仅解决同义词表、蕴涵表在表示概念及其概念之间关系上的局限性,而且降低人工智能对知识处理的复杂度,进而协助检索用户准确表达其内心信息需求,这是元搜索引擎研究的重要课题之一。

(2)检索词的“一词多义”和“多词一义”现象。“一词多义”和“多词一义”问题是影响元搜索引擎的查准率和查全率的重要影响因素。一方面,由于用户输入的关键词(Q)中缺少必要的上下文背景信息(C),很难避免自然语言的“一词多义”现象,影响查准率。例如在元搜索引擎 clusty. com[18] 中输入关键词“苹果”时,返回结果中与苹果电脑有关的信息和与水果类苹果有关的信息混合在一起。“一词多义”问题的主要原因在于用户所输入的关键词过于简单,缺乏必要上下文概念;另一方面,由于目前关键词检索仍然停留在字符匹配层次,尚未实现语义层次的匹配,无法解决自然语言中的“多词一义”现象,从而降低了查全率。例如,在元搜索引擎 clusty. com 中采用“电脑”或“计算机”作为关键词进行检索时,所返回的结果记录不同。

目前,解决“一词多义”的方法是“鼓励用户输入更详细的检索提问”,而解决“多词一义”的方法是“通过增设术语库同义词和近义词”。但是上述两种方法均存在一定的局限性。前者缺少用户检索提问的优化机制,后者缺少对概念之间的复杂语义关系,如继承、交叉的分析机制。因此,如何解决搜索提问中的“一词多义”和“多词一义”问题是提高跨领域海量信息资源元搜索引擎搜索效果的重要挑战之一。

(3)缺少自动调整搜索范围的弹性搜索机

制。从用户搜索行为习惯分析,元搜索引擎的返回结果不宜过多或过少。由于缺少自动调整检索范围的弹性机制,元搜索引擎往往返回过多的检索结果或过少的检索结果,导致用户无法一一浏览所有的返回结果或不能找出与需求有关的数据,影响元搜索引擎的查准率和查全率。

缺少自动调整搜索范围的弹性机制的主要原因有两个。一是缺乏自动扩展或缩小关键词语义范围的能力。目前,解决元搜索引擎的调整搜索范围功能主要通过改变搜索策略,如成员搜索引擎的选择、精确查询、高级查询或二次查询的方法实现。但是,这些方法不仅无法实现自动扩展或缩小的功能,而且仍然存在"一词多义"或"多词一义"的现象,限制了搜索引擎的查准率和查全率。二是缺少自动选择成员搜索引擎的机制。目前,元搜索引擎所采用的成员搜索引擎选择机制有两种。一种是基于系统选择,由系统决定选择哪几个成员搜索引擎,一般是具有一定知名度、使用率较高的主流搜索引擎,这是在元搜索引擎对成员搜索引擎的性能效率进行评价的基础上实现的;另一种是基于用户选择,在元搜索引擎的界面上提供了相应的选项供用户选择,在用户选项部分,用户能设定他们想使用的搜索引擎,并且可以随时改变这些设置[19]。从用户个性化信息搜索角度看,后者的搜索效率高于前者。但是,元搜索引擎尚未提供自动计算用户信息与成员搜索引擎相关性的机制,导致元搜索引擎无法实现成员搜索引擎的自动选择功能,影响了搜索查准率、查全率和查询速度。

(4)缺少通过用户反馈实现元搜索引擎增值的机制。目前,元搜索引擎中的用户研究主要侧重于个性化搜索服务[20],元搜索引擎只返回与用户提问相关性较大的结果记录集,而没有研究如何通过用户对搜索结果内容本身进行标注,实现元搜索引擎向知识引擎的转变。由于元搜索引擎中不仅缺少信息资源的自动爬取和本地存储索引库机制,而且还缺少对每次检索结果的本地存储机制,元搜索引擎的搜索效果过分依赖于各成员搜索引擎,并向各成员搜索引擎多次发送相同或相近的搜索提问,增加成员搜索引擎的负担,从而降低了元搜索引擎本身的搜索速度和稳定性。Web2.0 的成功经验证明,内容服务提供商的核心竞争力来自于重视创建和维护自己的核心竞争力数据资源库,尤其是带有用户评注的数据库[21]。因此,元搜索引擎应通过不断在本地积累搜索数据,并对其进行用户标注等增值操作,实现向知识引擎的转变。

3 方案设计

3.1 基本思路

针对上述研究挑战,结合跨领域海量信息资源搜索的实际需求,下文通过综合运用语义 Web 和 Web2.0 技术设计出面向跨领域海量信息资源的元搜索引擎,其基本思路如下:

(1)通过本体浏览和系统提示机制,协助用户准确表达其信息搜索意图,保证用户信息请求的准确表示。具体实现方法有三种:一是通过浏览领域本体,在领域本体中选择能够准确、规范表达其信息搜索意图的概念或实例作为其搜索关键词;二是当用户输入关键词时,元搜索引擎通过检索领域本体,将与用户输入的关键词相关的概念及其相互关系显示给用户,供用户进一步选择能够更准确表示个人信息需求的概念或实例;三是上述两种方法的整合与优化,即先采用第一种方法选择适当的关键词,然后采用第二种方法进一步改进或限定关键词,准确表示自己的搜索意图。

(2)利用领域本体对概念及其关系的规范化表示机制,帮助用户优化其检索词,从而解决关键词的"一词多义"和"多词一义"问题。一方面,采用用户关键词的上位概念、属性、下位概念或实例(表 1)补充用户输入的关键词,提高用户请求信息的详尽程度。上位概念、属性、下位概念或实例与关键词的相关程度较高,通过这些概念来扩展关键词可以较好地避免"一词多义"的问题;另一方面,通过关键词与领域本体中的概念之间的相似度计算得出与用户输入的关键词相关的概念或实例,克服目前的关键词字符匹配的检索策略,真正从语义上实现概念匹配,解决"多词一义"的问题。

表 1　采用 OWL 表示领域知识之间的关系

序号	相互关系	对应标记或属性
1	类之间的继承、等同、交叉关系	< rdfs:subclassOf > 、< owl:equivalentClass > 、< owl:disjointWith >
2	属性之间的继承、等同、互逆关系	< rdfs: subPropertyOf > 、< owl: equivalentProperty > 、< owl: inverseOf >
3	属性与类之间关系	< rdfs:domain > 、< rdfs:range >
4	类与实例之间关系	< rdf:Description > 、< rdf:type >
5	知识间的传递、对称、函数和反函数关系	owl: TransitiveProperty, owl: SymmetricProperty、owl: FunctionalProperty 和 owl:InverseFunctionalProperty
6	知识间的集合运算关系	< owl:unionOf > 、< owl:intersectionOf > 、< owl:complementOf >
7	同一个知识的不同版本关系	< owl:priorVersion >

资料来源: 朝乐门. 基于语义 Web 的知识处理流程及其技术框架研究[J]. 中国图书馆学报,2009(5):59 - 69.

(3)采用两种不同的方法解决自动调整搜索范围的弹性搜索问题。一是采用领域本体中的概念或实例之间的关系设计元搜索引擎的自动调整搜索范围的弹性机制。当搜索反馈结果较少时,元搜索引擎自动采用关键词的上位概念(或相关概念、下位概念、实例)进行扩展检索,找出更多的相关记录,实现元搜索引擎的自动扩展检索范围机制;当检索结果过多时,元搜索引擎自动采用下位概念(或属性限制、实例、规则推理)缩小检索范围,找出更准确的检索记录,实现元搜索引擎的自动缩小检索范围机制。二是采用三种不同的自动选择成员搜索引擎的方法。第一,通过用户个人信息(含用户基本信息和用户检索历史数据)与各成员搜索引擎记录之间相关性计算,选择相关度较大的搜索引擎作为成员搜索引擎;第二,通过用户输入的搜索提问与各成员搜索引擎记录之间相关性计算,选择相关度较大的搜索引擎作为成员搜索引擎;第三,根据搜索返回结果的多少来增加或减少成员搜索引擎的数量,支持自动扩展或缩小目标独立搜索个数。

(4)通过 Web2.0 用户标注机制实现元搜索引擎增值问题。为用户提供对检索结果进行相关性、可靠性评价和补充相关内容的标注功能,并把用户标注与搜索记录一起放在本地知识库中,以便在其他搜索中使用。随着用户标注数据的增多,元搜索引擎可以返回增值型搜索结果,实现元搜索引擎的增值服务。另外,系统通过自动捕获用户的搜索行为,如点击次数或阅读时间产生用户标注信息,并将其存储在本地知识库之中。用户标注机制不仅可以实现元搜索引擎的增值服务,而且还可以提高元搜索引擎的搜索返回速度。通过对每次搜索结果的用户标注及其本地化存储机制,在元搜索端不断积累知识,减少元搜索引擎对各成员搜索引擎的依赖程度,实现元搜索引擎向知识引擎的过渡。

3.2　框架设计

根据上述基本思路,本文对面向跨领域海量信息资源的元搜索引擎框架体系的设计可以分为用户、元搜索引擎、独立搜索引擎和海量信息资源四个不同层次。其中元搜索引擎层包括七个不同的组件,即用户代理、本体检索器、相关度计算器、独立搜索引擎代理、搜索结果与处理器、领域本体、本地知识库。相对于传统搜索引擎框架体系(图 1),本框架体系(图 2)的变化主要体现在三个方面:第一,采用用户代理、独立搜索引擎代理和搜索结果预处理器分别扩展了传统元搜索引擎中的检索请求代理、检索接口代理和结果显示代理;第二,新增了领域本体和本地知识库,用于存储用户评注过的搜索记录,实现元搜索引擎向知识库转变;第三,增设本体检索器和相关度计算器,用以确定用户输入关键词相关的概念,实现用户关键词的扩展与优化。

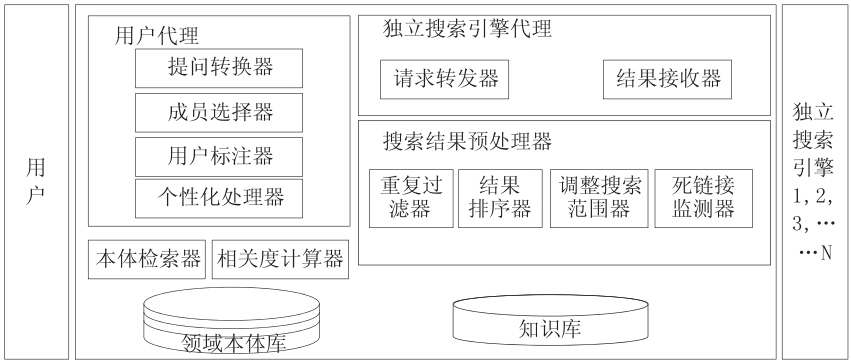

图 2　面向跨领域海量信息资源的元搜索引擎框架体系

(1)用户代理。用户代理的主要功能是帮助用户进一步明确表达其搜索意图,从而克服“用户准确表达自己的搜索意图困难”、“缺少根据用户需求自动选择成员搜索引擎机制”、“缺少搜索引擎与用户之间的交互机制研究”等问题。用户代理由提问转换器、成员选择器、用户标注器和个性化处理器四部分组成。其中,提问转换器的功能是实现用户搜索意图(P)到搜索提问(Q)的转换;成员选择器的功能是根据用户搜索提问(Q)确定一次搜索所涉及的目标搜索引擎;用户标注器为用户对搜索反馈结果进行标注提供支持;个性化处理器是实现跨领域海量搜索引擎的个性化处理功能。

(2)独立搜索引擎代理。独立搜索引擎代理的主要作用是转发用户搜索提问(Q)和接收独立搜索引擎所反馈的结果记录集。其中,请求转发器的功能是接收用户代理发送的搜索提问(Q),并转发给所对应的独立搜索引擎;结果接收器的功能是接收独立搜索引擎返回的结果,并转发给系统的另一个组件——“搜索结果预处理器”。独立搜索引擎代理从每一条返回结果记录中自动抽取多个特征关键词,并调用“相关度计算器”计算独立搜索引擎返回的搜索结果与用户请求之间的相关度,并参照一定的阈值,过滤相关度较低的返回结果,并将相关性较高的检索结果集转发给“搜索结果预处理器”。

(3)搜索结果预处理器。其主要功能是对独立搜索引擎代理返回的结果进行预处理,具体包括重复过滤器、结果排序器、调整搜索范围器和死链接监测器。其中,重复过滤器的作用是对各独立搜索引擎代理所返回的搜索结果进行重复过滤,实现访问负载均衡,排除重复信息;结果排序器的功能是采用一定的排序策略对各独立搜索引擎代理返回的结果进行有效排序,以便用户的阅读和统计分析;调整搜索范围器的作用是当各独立搜索引擎代理返回的搜索结果记录过少(或过多)时,采用一定策略进行扩展(或缩减)搜索范围,保证搜索结果记录集的数量,它是克服目前搜索引擎中普遍存在的“缺少自动调整搜索范围的弹性机制”这一问题的重要部件;死链接监测器的功能是判断目标记录是否存在,是否能被用户访问,以避免用户无法打开搜索引擎所返回的结果记录中的超链接这一情况的发生。

(4)领域本体库与知识库。领域本体库中存放多个领域的本体,是计算概念之间的相似度、提问转换、弹性搜索的重要依据。领域本体相对稳定,一般由知识链头部——领域专家负责建立,并由知识链尾部——长尾用户负责管理和维护。领域本体库中的本体之间可以存在继承关系,以表示大领域和小领域的知识建模;本框架体系中的知识库是指领域本体库的实例化库,用以存放用户标注过的搜索结果集,是解决目前搜索引擎所面临的“缺少元搜索引擎向知识引擎转换机制”这一问题的主要部件。一般情况下,这种知识库具有相对动态的特征,应由知识链长尾部分的搜索用户负责建立,并由知识链头部用户即领域专家负责定期管理。

(5)本体检索器和相关度计算器。本体检索器的功能是通过调用相似度计算器对领域本体库进行检索,返回相关概念,如同等概念、上位概念、下位概念以及对应实例等,是解决目前

搜索引擎中存在的“一词多义”和“多词一义”问题的重要部件。

3.3　流程设计

面向跨领域海量信息资源的元搜索引擎框架(图2)中的各组件之间的交互关系如图3所示。本文提出的面向跨领域海量信息资源的元搜索引擎的使用流程如下:

(1)使用本搜索引擎的用户需要在系统中填写自己的两种基本信息,即能力领域信息和兴趣领域信息。其中,能力领域信息代表用户的现有知识水平和需求,兴趣领域信息代表用户的未来知识需求。用户信息均以OWL的形式存放在搜索引擎本地知识库中,其主要作用有三个:第一,个性化推送。系统利用用户兴趣信息,进行自动扩展用户提问、成员搜索引擎的选择、返回结果处理等个性化服务;第二,社会网络分析。系统利用用户能力信息和兴趣信息,进行深度网络分析,实现信息协同过滤和搜索引擎与人肉搜索的集成;第三,确定领域专家和领域初学者。

(2)用户输入关键词时,系统通过自己的本体检索器对用户关键词与多领域本体库中的概念进行相似度计算,返回与关键词相关的“本体片段”。

(3)用户在系统返回的本体片段中选择能够准确表达其信息需求(R)的概念结点(N)或结点集合(S)。

(4)系统采用概念结点(N)的上下文扩展用户检索词,计算用户提问与各独立搜索引擎之间的相似度,选择目标成员搜索引擎,并根据不同成员的调用接口进行必要的提问转换工作,将用户搜索提问(Q)转发给各成员独立搜索引擎。

(5)搜索结果预处理器对各成员搜索引擎返回的结果进行重复过滤和相关排序。如果搜索结果过多或过少,系统将自动调整搜索范围,重新搜索。

(6)搜索结果预处理器将重复过滤和相关排序后的搜索结果返回给人机界面。

(7)用户对系统返回结果进行标注,标注数据放在本地知识库中,以便于搜索结果的不断改进。此外,系统可以自动捕获用户的搜索行为数据,如某一条记录被点击次数或用户在某条记录上停留的时间。

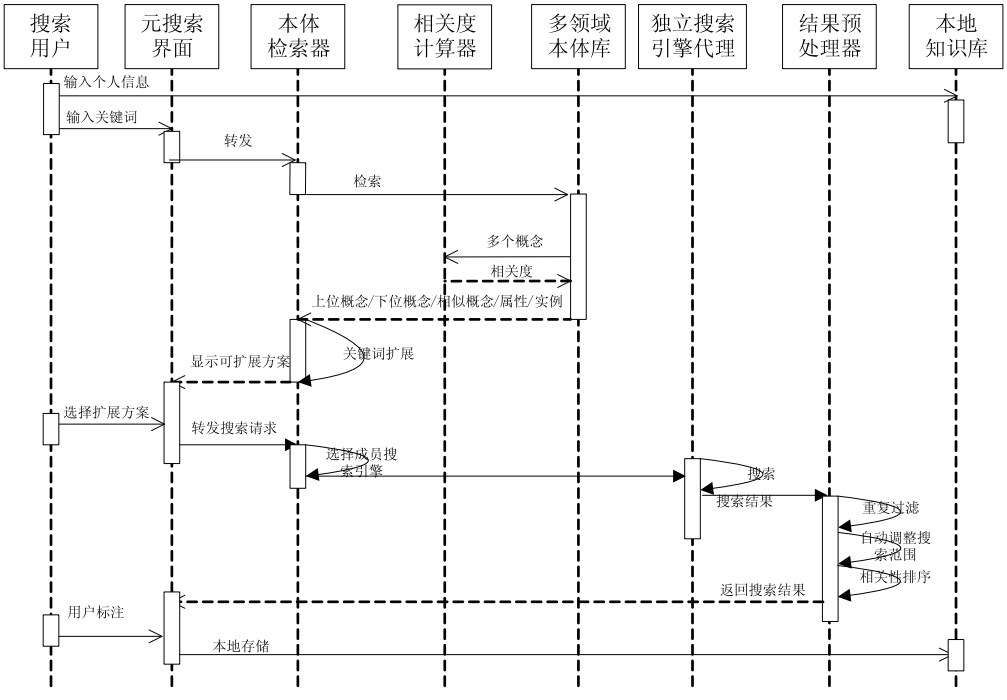

图3　面向跨领域海量信息资源的元搜索引擎各组件间交互关系

随着搜索系统的不断使用,上述过程的不断重复,本地搜索引擎不断积累带有用户标注的搜索结果,实现从元搜索引擎向知识检索系统的转变。因此,本搜索系统可以实现用户越多、使用次数越多,搜索效果越好的良性发展。

3.4 关键技术

从上述框架设计和流程设计可以看出,本文提出的面向跨领域海量信息资源的元搜索引擎的关键技术有八种。

(1)提问转换技术。从跨领域海量数据资源搜索体系结构(图2)和内部交互(图3)可以看出,提问转换是用户代理的重要功能之一。提问转换技术的应用如下:首先,根据用户输入的关键词(Q1),在多本体或跨领域本体库中进行关键词检索,定位查询词在领域本体中对应节点(Q2)的位置,并进一步计算对应的相关节点(Q3),如父节点、子节点和兄弟节点;其次,根据用户记录信息对结果节点集 Q3 进行过滤,得出个性化节点集(Q4),并将个性化结果节点集 Q4 反馈至用户界面;接着,用户可以在个性化结果节点集 Q4 中选择所需要的"领域"(D1,D2,D3,..),选择能够较好地表达个人信息搜索意图的词语(Q5),进一步明确表示自己的检索提问(Q6),实现提问转换,得出搜索提问(Q7);最后,当独立搜索引擎的检索接口要求与元搜索引擎不同时,需要将提问转换为独立搜索引擎的提问 $Q8_i$(其中下标 i 代表的是第 i 个独立搜索引擎)。可见,在跨领域海量数据资源搜索过程中存在多种提问转换工作,不同类型的搜索引擎所采用的关键技术有所不同。

(2)相似度计算和语义距离计算方法。相似度计算是指计算两个概念之间的相似度。由于领域本体和知识库中的概念之间关系的形式化描述,本文提出的跨领域海量数据资源搜索机制中的相似度计算的效果好于传统搜索机制。基于本体的相似度计算和语义距离计算是目前本体及其应用研究中的热点问题,从总体上看,基于本体的语义相似度计算方法划分为基于距离的语义相似度计算(Edge Counting Measures)、基于内容的语义相似度计算(Information Content Measures)、基于属性的语义相似度计算(Feature-based Measures)和混合式语义相似度计算(Hybrid Measures)四大类型[22]。

(3)弹性搜索技术。弹性搜索技术是指元搜索引擎可以根据各搜索引擎反馈结果值的多少自动改变检索策略。当检索结果过少,则用上位词进行扩展检索,否则通过下位词进行精准检索。另外,可以采用相似度计算技术,计算关键词与用户查询词之间的相关度高低,并采用相关度较低的术语进行提问扩展。

(4)用户标注技术。用户标注技术可以分为两大类,即直接标注和间接标注。直接标注是指用户对搜索结果进行有意识的标注行为,具体表现形式为用户评价、用户补充和用户打分;间接标注是指搜索引擎系统自动记录用户行为,如停留时间、打开次数、下载次数等。日志分析技术是最典型的间接标注技术之一。

(5)用户搜索行为特征建模。用户搜索行为特征建模的数据主要源自三个方面,即用户填写的个人信息、用户行为记录挖掘、用户标注信息分析以及用户社会网络分析。为使用户信息与领域本体进行匹配,以便进行提问转换、相似度计算、精准检索、个性化推送等服务,面向跨领域海量信息资源的元搜索引擎需要采用语义 Web 技术进行用户搜索行为特征建模。FOAF(Friend of a Friend)采用 RDF 技术记录个人信息(如 foaf:Agent、foaf:Document、foaf:Group 等)和个人之间的关联信息(如 foaf:knows、foaf:made/foaf:maker、foaf:fundBy、foaf:menber 等),并在个人信息之间建立计算机可理解的互连关系,强调了人在知识联网中的重要作用[23-25],是典型的采用语义 Web 技术进行用户信息建模的关键技术之一。

(6)个性化搜索技术。个性化搜索体现在三个方面,即目标搜索引擎的选择、提问转换和搜索返回结果的处理。目标搜索引擎的选择可以通过计算用户信息向量空间与搜索引擎信息向量空间的相似度,将相似度较高的搜索引擎作为目标搜索引擎;搜索返回结果的处理可以采用搜索反馈结果与用户信息进行相似度计算或语义距离计算,向用户提供个性化程度较高

的检索服务。

(7)大规模人机协同技术。基于语义 Web 的大规模人机协同技术是未来知识处理技术的重要发展趋势,也是本文所提出的面向跨领域海量信息资源的元搜索引擎的重要特征之一。该搜索引擎鼓励来自多个不同领域的长尾用户采用领域本体和知识库的建设活动,从而充分发挥人与计算机在知识处理中的互补性优势。语义 Web 和 Web2.0 技术以及二者的结合是新兴的大规模人机协同技术。根据 Tim Berners-Lee 的观点,语义 Web 技术分层体系结构自下而上依次为 Unicode 和 URI 层、XML、RDF(S)、本体、逻辑、证明和信任,不同层次之间遵守两个重要原则:一是向下具有兼容性,即位于某一层的代理应能够解释和适用底层的信息;二是向上具有可理解性,即位于某一层的代理能部分地使用更高层次的信息[26]。语义 Web 技术的面向计算机可理解的知识表示以及语义 Web 知识处理流程中对前端控制的重视,降低了计算机知识处理的难度,使现有的知识工程领域的研究成果基本满足计算机知识处理的需要,为人和计算机协同完成知识处理提供思路和平台[27]。Web2.0 技术包括 Blog、RSS、Wiki、Tag、SNS、P2P 和 IM 等,基于 Web2.0 的大规模人机协同知识处理具有三大特征:在推动非组织知识转化为组织知识的过程中,组织知识管理的范围必须延伸至知识链的"长尾",而不能仅仅停留在其"头部";在深度挖掘和充分利用组织已有知识的过程中,应强调基于知识库管理和挖掘的核心竞争能力,而不是来自基于软硬件设施的核心竞争力;在创建组织知识生态系统时,必须强调知识管理中的平等、协同、自组织,而不是强制、统治、控制与高度集中管理[28]。可见,Web2.0 技术降低了知识链头部用户在知识处理中的高成本,为大规模面向跨领域海量信息资源的元搜索引擎提供了新的视角和技术手段。另外,近年来出现了语义 Web 与 Web2.0 融合的趋势[29-31],更好地推动了大规模人机协同的知识处理。

(8)元搜索引擎的增值技术。推动元搜索引擎向增值型搜索引擎的转变,进而实现建设区别于 Google 等以内容为中心的传统搜索引擎的、带有用户标注的、以搜索用户为中心的新一代增值型搜索引擎。随着跨领域海量信息资源的元搜索引擎向增值型搜索引擎的转变,系统的查全率和查准率将得到不断提高,查询速度将得到较大改善。因此,本文所提出的跨领域海量信息资源的搜索机制的效率与用户的使用直接相关,用户数量越多,用户标注的数据越多,系统的搜索效率越高。在元搜索引擎向增值型搜索引擎的转变过程中,领域本体技术和知识库技术是两项关键性技术。

4　结论

为了论证本文研究结论的合理性,我们开发了演示验证原型系统(图 4)。该系统以笔者开发的咨询公司知识地图系统[32]为基础,采用 OWL 领域本体测试数据集和 Jena 接口[33]对其高级检索功能进行了一定的改进。作为演示验证系统,目前主要完成了如下五项工作:第一,采用 Protégé[34]为编辑工具在领域本体库中分别建立了医学和计算机两个不同领域的测试本体及其实例化测试数据;第二,采用 Java 语言、Jena 接口和 SPARQL[35]语法开发了简单的本体检索器,可以计算出用户关键词的上位概念、下位概念、同等概念和实例概念,组成相关概念集;第三,利用文献[36]中讨论的计算本体语义距离算法,开发了简单的语义距离计算器,模拟了相关度计算器的功能;第四,将本体检索器返回的相关概念集以下列列表框的形式显示给用户;第五,实现了检索结果的用户标注功能,包括相关性打分、补充信息和评价信息。改进后的应用方法如下:用户在"高级检索"界面的"领域范围"中可以同时选择多个领域,并在"搜索提问"中输入查询关键词,系统在对应领域本体库中检索相关概念,进行提问转换工作,以下拉列表的形式显示给用户,用户可以选择所需的搜索表达式,实现准确表达个人信息搜索意图的目的,从而提高系统的搜索效果。以搜索专家"王林的联系方式"为例,用户可以选择计算机科学、医学和社会学等多个领域,当输入姓名

“王林”时,系统自动提示三个搜索提问,分别为“[医学]…[内科][王林][联系方式]”、“[医学]…[外科][王林][联系方式]”、“[计算机]…[OS][王林][操作系统教程]”,用户可以根据自己的需要选择搜索提问,得出准确的查询结果。

图 4 演示验证系统

总之,本文以跨学科学术研究实际需求为驱动,针对基于元搜索引擎的跨领域信息资源搜索所面临的四个关键问题,提出了面向跨领域海量信息资源的元搜索引擎。在该元搜索机制中,通过本体浏览和系统提示机制,协助用户准确表达其信息搜索意图,保证用户信息请求的准确表示;利用领域本体对概念及其关系的规范化表示机制,帮助用户优化其检索词,从而解决关键词的“一词多义”和“多词一义”问题;采用领域本体中的概念或实例之间的关系,设计元搜索引擎的自动调整搜索范围的弹性机制;采用三种不同的自动选择成员搜索引擎的方法;通过 Web2.0 用户标注机制实现元搜索引擎的增值服务;通过对每次搜索结果的用户标注及其本地化存储机制,在元搜索引擎中不断积累知识,减少元搜索引擎对各成员搜索引擎的依赖程度,实现元搜索引擎向知识引擎的过渡。

参考文献:

[1] Sullivan D. Search engine coverage study published[R/OL]. [2010-12-20]. http://searchenginewatch.com/2167411.

[2] Spink A, Jansen B J, Blakely C, et al. A study of results overlap and uniqueness among major Web search engines[J]. Information Processing and Management, 2006, 42(5):1379-1391.

[3] Bazac D. Features - The meta search engines: A web searcher's best friends [R/OL]. [2010-05-16]. http://www.llrx.com/features/metasearch.htm.

[4] 原福永,梁顺攀. 元搜索引擎的现状与发展[J]. 计算机工程与设计, 2005(12):3278-3280.

[5] 孟晓明. 元搜索引擎及其发展[J]. 中国信息导报,2007(3):56-59.

[6] 种梅,刘方爱. 元搜索引擎中的成员选择和结果合并策略研究[J]. 计算机工程与设计,2007(21):5125-5127.

[7] 曹林,韩立新,吴胜利. 元搜索引擎排序技术综述[J]. 计算机应用研究,2009(2):411-414.

[8] Rao B S, Rao S V, Sajith G. A user-profile assisted meta search engine[C]. Conference on Convergent Technologies for Asia-Pacific Region, TENCON, 2003: 713-717.

[9] Davison B D. The potential of the metasearch engine[C]. ASIST 2004: Proceedings of the 67th ASIST Annual Meeting, 2004 (41):393-402.

[10] Sadeghi H. Assessing metasearch engine performance[J]. Online Information Review, 2009, 33(6):1058-1065.

[11] 朱晓丽. 中外九大元搜索引擎的比较研究[J].

数字图书馆论坛,2007(9):57-62,67.
[12] 王艳芬,杨东东,王琼. 基于本体的元搜索引擎的设计与实现[J]. 计算机工程与设计,2008,29(13):3522-3525.
[13] 王春云,秦杰,胡双双. 基于本体的元搜索引擎技术研究[J]. 微型电脑应用,2008(9):8-9.
[14] 沈宇,黄卫东. 基于领域本体的元搜索技术研究[J]. 信息通信,2008(2):17-20,39.
[15] Moskovitch R, Shahar Y. Vaidurya: A multiple-ontology, concept-based, context- sensitive clinical-guideline search engine [J]. Journal of Biomedical Informatics, 2009,42(1): 11-21.
[16] Kanteev M, Minakov I, Rzevski G, et al. Multi-agent meta-search engine based on domain ontology[C]. Autonomous Intelligent Systems: Agents and Data Mining: Proceedings Second International Workshop, AIS-ADM 2007,269-274.
[17] 黄崑,赖茂生. Web 信息检索技术及研究进展[J]. 现代图书情报技术,2004(5):44-57.
[18] Vivísimo. Clusty[R/OL]. [2010-05-16]. http://clusty. com/.
[19] 任洪平. 中文元搜索引擎成员搜索引擎的选择策略研究[J]. 图书馆学研究,2009(1): 40-43.
[20] 朱前东. 搜索引擎个性化检索研究综述[J]. 图书馆学刊,2008(6):14-17.
[21] O'Reilly T. What is Web 2.0: Design patterns and business models for the next generation of software' O'Reilly[R/OL]. [2009-08-27]. http://www. oreillynet. com/lpt/a/6228.
[22] 孙海霞,钱庆,成颖. 基于本体的语义相似度计算方法研究综述[J]. 现代图书情报技术,2010(1):51-56.
[23] Brickley D, Miller L. FOAF vocabulary specification 0.91 namespace document 2 November 2007 - OpenID Edition [EB/OL]. [2009-04-05]. http://xmlns. com/foaf/spec/2007.
[24] Dumbill E . XML Watch: Finding friends with XML and RDF [EB/OL]. [2009-04-05]. http://www. ibm. com/developerworks/xml/library/x-foaf. html
[25] FOAF Project . FOAF Project Homepage [EB/OL]. [2009-04-25]. http://rdfweb. org/foaf/.
[26] Berners-Lee T, Hendler J, Lassila O. The Semantic Web [J]. Scientific American, 2001, 285(5):34-43.
[27] 朝乐门. 基于语义 Web 的人机协同知识处理研究[J]. 图书情报工作,2009(24):115-119.
[28] 朝乐门. Web2.0 在组织知识管理中的应用研究[J]. 情报资料工作,2010(2):49-52.
[29] Greaves M, Mika P. Semantic Web and Web 2.0 [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(1):1-3.
[30] Hendler J, Golbeck J. Metcalfe's law, Web 2.0, and the Semantic Web [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(1):14-20.
[31] Ankolekar A, Krotzsch M, Tran T, et al. The two cultures: Mashing up Web 2.0 and the Semantic Web[J], Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(1): 70-75.
[32] 朝乐门. 咨询公司知识地图系统的研究与开发[J]. 图书情报工作,2009(4):61-64.
[33] Jena Semantic Web Framework. Jena - A semantic Web framework for Java[OL]. [2010-07-21]. http://jena. sourceforge. net/.
[34] Protégé. The protégé ontology editor and knowledge acquisition system [OL]. [2010-07-22]. http://protege. stanford. edu/.
[35] W3C. SPARQL query language for RDF[OL]. [2010-08-01]. http://www. w3. org/TR/rdf-sparql-query/.
[36] Tsang V , Stevenson S . A Graph - Theoretic framework for semantic distance[J]. Computational Linguistics,2010, 36(1):31-69.

朝乐门 北京石油化工学院经济管理学院讲师,清华大学计算机系在站博士后。通讯地址:北京清华大学信息技术研究院 FIT 楼 1-311 室。邮编:100084。

张 勇 清华大学信息技术研究院副研究员,硕士生导师。通讯地址同上。

邢春晓 清华大学信息技术研究院研究员,博士生导师。通讯地址同上。

(收稿日期:2010-07-16;修回日期:2010-08-05)