基于本体的元搜索引擎技术研究①

Research of Ontology Based on Meta Search Engine

王春云　秦　杰　胡双双　（河南工业大学 信息科学与技术学院 河南 郑州 450001）

摘　要: 针对现有搜索引擎的查询结果相关性低和缺少语义理解能力等问题,提出了一种基于本体的元搜索引擎模型。主要应用基于本体的用户个性模型和本体语义分析关联方法来提高元搜索引擎的查询效率。目的通过领域本体的语义理解应用,为用户提供查询意图个性化的有效推测和关键词本体的查询优化,通过验证表明,该搜索模型实现了查询结果的有效优化。

关键词: 语义 本体 关联 用户个性化模型 元搜索引擎

1　引言

当今,在如此浩瀚的网络资源中如何有效地获取所需信息成为人们关心的话题。baidu,google 这些搜索工具的出现与发展为用户查询信息带来了方便,然而任何单个搜索引擎都只能涵盖 WWW 上很少一部分资源。元搜索引擎的出现在一定程序上解决了这些问题,它融合多个搜索引擎的检索结果,扩大了检索范围[1]。

但是现有的搜索引擎由于缺乏语义理解能力,查询的结果存在大量垃圾信息。随着本体(Ontology)和语义网(Semantic Web)技术的不断发展和应用,有关基于语义本体的检索系统成为研究热点[2-4]。本文结合语义本体和搜索引擎技术提出了一种基于本体的元搜索引擎系统模型,利用个性化用户模型和本体处理等技术,对用户的兴趣和查询习惯进行挖掘分析,并对查询关键词进行语义本体分析推理,来有效理解用户意图,提高查询结果的相关性。

2　基于本体的元搜索引擎模型

元搜索引擎是一种建立在搜索引擎之上的搜索引擎,它调用其他独立搜索引擎(也称为源搜索引擎),将用户的查询请求经过一定的处理后提交给其他的源搜索引擎,然后收集各个源搜索引擎的结果并经过相关处理后按相关性顺序返回前端用户。元搜索引擎一般包括四个方面[5]:用户查询请求预处理模块,查询请求分发模块,结果加工模块,页面显示模块。本文设计了一种基于本体的元搜索引擎模型如图1:

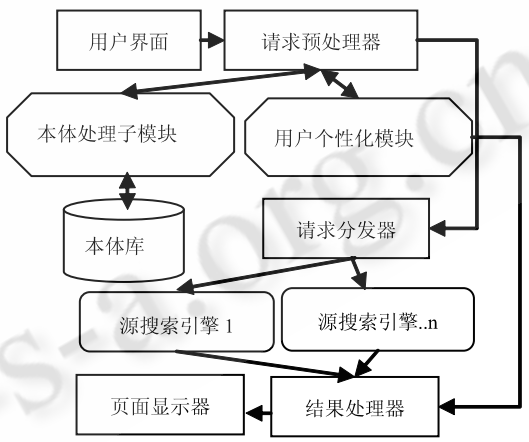

图1　基于本体的元搜索引擎模型

2.1 图形化用户界面

用户界面是用户和搜索引擎进行交互的途径,简洁实用的界面是最佳选择。本文将在用户界面中增加查询所属领域的选择,以便在以后的模块中进行相关领域内的匹配。

2.2 查询请求预处理

查询预处理模块是该基于本体的元搜索引擎系统的关键模块,它主要包括用户意图的个性化分析,关键字的语义分析、本体库的映射等等。

① 基金支持:河南工业大学科研项目(网络环境下半结构数据存储与查询技术研究),编号:2006BS009

2.2.1 用户意图个性化模型

为了提高对用户意图理解能力和用户的查询个性化，本文借鉴文献[6]建立基于本体的用户个性化兴趣本体库如图 2 所示。

在该用户模型中用户行为分为显式和隐式两种，显式行为主要通过用户提供的背景知识感兴趣的领域和主动反馈的信息，隐式行为则是通过用户的浏览内容等表现出来，其中隐含了用户的个人兴趣和偏好。基于本体的个性化用户模型定义为一个三元组：

UserModel = (PersonalI, PersonalO, PersonalR)

其中，PersonalI 代表用户个人信息，包括用户姓名、性别、年龄等基本信息和学历、专业、兴趣描述等背景知识；PersonalO 是一个包含了用户信息的个性化领域本体；PersonalR 是用户的个性化信息需求。

本系统中根据上述定义建立一个简单的用于实验的用户兴趣库，包括用户主动反馈和访问日志关键词构成的个性化信息。

当用户登陆并输入查询关键词时，查询预处理器将根据用户信息来匹配用户兴趣本体库，选择或替换成更接近用户的相关查询领域内的查询关键词(组)。

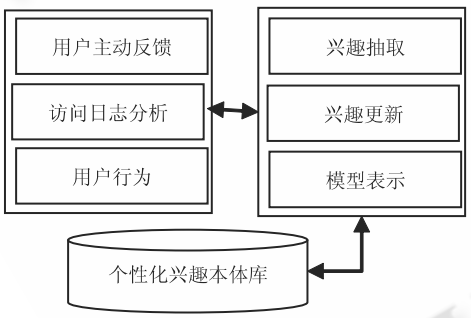

图 2　基于本体的个性化用户模型

2.2.2 语义本体匹配(关联)

针对基于语义关联的查询优化问题本文则采用文献[7]提出的方法。该方法的基本思想是：综合词法关系和词义分析的查询优化方法，通过对查询关键字词法特性和本体实例之间语义关联强弱的分析，提高查询关键字到本体概念映射的完整性和准确率。例如输入 university，按照传统的关键字匹配的搜索方法，用 college、academy 等描述大学的文档都会被遗漏。使用 WordNet 能知道 college、academy 与 university 相关，但其他一些与“大学”无关的词语，如 jury(陪审团)在 WordNet 中也是与 university 相关的。如果简单地使用相关概念进行匹配搜索，虽然能在很大程度上解决信息遗漏，但同时将会带来更多的无用数据。

本文对其进行改进：结合基于本体的个性化用户模型进行用户输入查询内容的优化。算法主要步骤如下：

(1) 接受用户输入，将关键词保存在集合 $K = \{k_1, k_2, \cdots, k_n\}$，$k_i$ 为用户输入的第 i 个关键词，n 为关键词个数。

(2) 对关键词集合进行用户个性化模型处理后，处理过程中将根据用户的兴趣本体库对关键词进行相对领域内的替换、增删关键词等操作。得到关键词集合 $N = \{n_1, n_2, \cdots, n_k\}$。

(3) 使用 WordNet 进行词法处理，得到每个关键词 ni($1 \leqslant i \leqslant k$)的在领域内的同义词集合 $Si = \{S_{i1}, S_{i2}, \cdots, S_{im}\}$，m 为 ni 的同义词个数，$n_i \in S_i$. 记元素 $S_{ij} \in S_i$ 的词法相关度为 R_{ij}，表示 S_{ij} 和 ni 的词法关联程度，$0 \leqslant R_{ij} \leqslant 1$，ni 与自身的词法相关度为 1。

(4) 将领域内同义词集合中的元素映射到本体库，本体集合 $Ti = \{T_{i1}, T_{i2}, \cdots, T_{if}\}$ 保存 Si 中词语对应的本体实例，共得到 k 个本体集合。记元素 $T_{ij} \in T_i$ 的词法相关度为 R'_{ij}，等于映射到它的 S_i 中元素的词法相关度。

(5) 设集合 $W = \{(a_1, a_2, \cdots, a_k) \mid (a_i \in T_i, T_i \neq \Phi, 1 \leqslant i \leqslant k) \cup (a_i = \beta, T_i = \Phi, 1 \leqslant i \leqslant k)\}$，Φ 表示空集，计算 $w_n \in W$ 的语义关联值 V_n。

$$V_n = \alpha \sum_{i,j=1}^{k} v_{ij} + \mu \sum_{i=1}^{k} r_i$$

式中，v_{ij} 为 w_n 中元素 a_i、a_j 的语义相关度；r_i 为元素 ai 的词法相关度；β 为用户输入关键词中没有找到同义匹配的关键词；α、μ 分别是语义相关度和词法相关度的权重系数，$\alpha \geqslant 0$、$\mu \leqslant 1$，$\alpha + \mu = 1$。综合考虑词法和语义的关联强弱。

(6) 将集合 W 中元素按语义关联排序，取语义关联程度最高，即 V 值最小的 κ 个元素。对于每个 wi($1 \leqslant i \leqslant \kappa$)，用知识库中对本体 $a_i \in w_i$ ($a_i \neq \beta$) 的概念描述代替关键词集合 N 中的第 j 个元素，向用户返回得到 κ 个优化建议，并结合用户个性化把其感兴趣的建议排到前面，以供用户根据自己情况更好地选择。

改进后的算法加入了用户个性化本体库，在进行语义同义词匹配之前先根据用户本体兴趣库确定其查

询领域,进行相关领域内关键词的替换。这可以减少了无关领域内同义词的本体语义匹配所消耗的时间和资源。

2.3 查询分发器

查询分发器的功能是将用户的查询请求根据知识库生成搜索调度策略并转化为各源搜索引擎的连接请求。调度策略借鉴文献[8]采用数据挖掘技术,根据源搜索引擎工作情况的记录,用决策树归纳分类技术生成元搜索引擎调用策略。它首先学习各个源搜索引擎在某条件下(如搜索主题)和查准率情况。然后进行分类。当用户有搜索请求时,会根据条件预测哪些搜索引擎可以满足用户查找精确信息的要求。同时建立评估度量,给出源搜索引擎的查准率,查全率的综合评估来确定兴趣度阈值。在建立调用策略时,不满足阈值的策略被认为是不感兴趣的。

2.4 结果处理器

结果处理器的功能是收集各个源搜索引擎的检索结果,按一定的合成算法处理后返回给用户。借鉴文献[9]提出以 Web 网页内容特征库为基础实现对查询短语进行语义理解的方法,加入二次检索模块,它的功能是接收各源搜索引擎的返回网页并过滤掉相同的网页,然后利用面向 Web 网页内容的特征库理解查询要求,对网页重新检索,计算网页与用户查询请求的相关度并对网页排序,最后交给用户查看。

3 技术实现及结果分析

3.1 实现环境

本文基于 Java 语言建立一个以研究为目的元搜索引擎模型。开发工具主要是 MyEclipse 6.0。本体技术方面利用 Protégé 3.4[10] 构建一个简单的本体库,存为 owl 文件的形式,在 MyEclipse 6.0 中加载 jena[12] 2.5.5,然后通过 jena API 加载构建的本体模型,使用 jena 执行 sparql 查询。

3.2 结果分析

下面以一个查询为例,验证本系统的有效性。

假设用户是不常利用搜索工具,对计算机领域知识了解不多。其目的是想查询有关论文的情况,但是只知道一些大众了解的计算机词汇。

假如在查询界面输入计算机领域常用词汇“网络”,在查询预处理器中进行处理,根据个人主动反馈建立本体兴趣记录和关键词本体查询,把关键字“网络”优化为“计算机网络”,然后匹配到“计算机网络基础教程”和“计算机类期刊”的结果供选择,根据查询目的选中“计算机类期刊”作为查询的关键字。可以看出实验选取的关键字可以正确地进行优化。

然后,用关键字“网络”在 baidu 中进行搜索,结果如下:“百度一下,找到相关网页约 100,000,000 篇”(注 2008 – 4 – 26 日查询)。其中信息包括计算机网络、网络电视、网络小说、网络游戏、网络广告、公司名字含有“网络”二字的公司信息、更有甚者是在网络上传播的不健康信息,等等。

利用“计算机类期刊”在 baidu 进行搜索,共搜到 216,000 条信息。通过分析可以看出显示的结果含有大量重复信息和无关信息,而其中前 10 页的结果中基本上涵盖了所需的 80% 以上的信息。因此本查询模型中采用提取源搜索引擎前 100 条记录,然后并对其中重复的结果进行处理。在模型中查询得到 192 条记录,浏览结果后发现基本上能满足查询需求。但是本查询系统有一个的缺点,那就是的查询速度比 Baidu 的查询速度慢,且构建的本体还只是以研究为目的一个小的模型。

4 结束语

语义本体和元搜索引擎相关技术的出现与发展,为解决现有搜索工具的查询覆盖率低、结果相关性差等问题提供了支持。本文借鉴现有技术提出了一种基于语义的元搜索引擎模型系统。通过领域本体的语义理解,为用户提供个性化和查询信息的语义扩展,意在为实现元搜索引擎更好的查全率和查准率用出贡献。但针对结构化更完善的本体库的建立和语义关联算法性能的提高,还需要更多人员投入更大的研究力度。今后将会在现有工作的基础上,对本体库,本体语义关联等方面做进一步的研究。

参考文献

1 张俭恭,陈定权,吴振新. 关于搜索引擎与元搜索引擎的讨论. 现代图书情报技术,2002(2):36 – 38.

2 Wollersheim D, Rahayu W. Ontology based query expansion framework for use in medical information systems. International Journal of Web (下转第 95 页)

(上接第 98 页)

Information System, 2005, 1(2): 101 – 115.

3 Fiaidhi J, Mohammed S, Jaam J, et al. A stardard frameword for personalization vis ontology – based query expansion. Pakistan Journal of Information and Technology, 2003, 2(2): 96 – 103.

4 Li Wenjie, Feng Zhiyong, Li Yong, et al. Ontology – Based Intelligent Information Retrieval System CCECE 2004 – CCGEI 2004. 2004: 373 – 376.

5 郭少友. 元搜索引擎的原理与设计. 情报科学, 2005 (2): 245 – 246.

6 陈琳娜. 基于本体的个性化用户模型研究: [硕士学位论文]. 秦皇岛: 燕山大学信息科学与工程学院, 2006.

7 梅翔, 孟祥武, 陈俊亮, 徐萌. 一种基于语义关联的查询优化方法. 北京邮电大学学报, 2006, (12): 108 – 109.

8 刘丽, 孙燕唐. 智能元搜索引擎的设计与实现. 计算机工程, 2003, 29(6): 119 – 120.

9 曹二堂. 一种基于语义理解的元搜索引擎的研究. 计算机工程, 2006, 32(7): 210 – 211.

10 Protégé User documentation. [Online]. Available; http://protege.stanford.edu/doc/users.html/.

11 Jena Semantic web Toolkit, [Online]. Available; Http://Jena.sourceforge.net/.