

一个基于本体的信息检索代理的设计

杜文华

(中南民族大学管理学院 武汉 430074)

摘 要 描述了一个基于本体的信息检索代理,阐述了各部分的功能,具体介绍了其主要的四个问题:领域本体的构造、三个抽象层的设计、查询模型的使用以及聚合函数的定义。

关键词 本体 信息检索代理 概念查询

目前在 Internet 上存在大量的通用搜索引擎,如有名的 google、百度,用于检索满足某种条件的 Web 网页;也有一些用来检索参考文献的专业引擎,如 PubMed 是检索 MedLine 数据库的。

由于编目方法或检索环境不了解,很多用户觉得明确地表达设计好的检索要求很困难。事实上,正如使用 Web 搜索引擎一样,为了进行有效的检索,用户可能需要花费大量时间来组织他们的检索条件。通常的情况是,用户先输入条件进行查询,看是否检索到信息,检索到的信息是否满足他的需求。大多数时候用户得到大量的、他难以处理的文档,其余时候,他将查询条件限制得太严格,以至于查不到有效信息。这时他就不得不重新组织查询。

针对这种情况,本文设计了一个基于本体的信息检索代理,来解决专业领域典型的检索问题。为了实现这一目标,设计了不同的模块来产生查询,评价结果,如果有必要,重新设计查询,最后将结果显示给用户。第一节是对整个系统的概述,第二节是解释专业数据库需要哪种类型的信息检索,第三节描述创建本体的过程,第四节解决查询模式设计问题,第五、六节介绍查询的产生及评价,最后给出了结论和将来要深入的工作。

1 概述

下面给出了本文设计的基于本体的信息检索代理模板的一个通用方案。

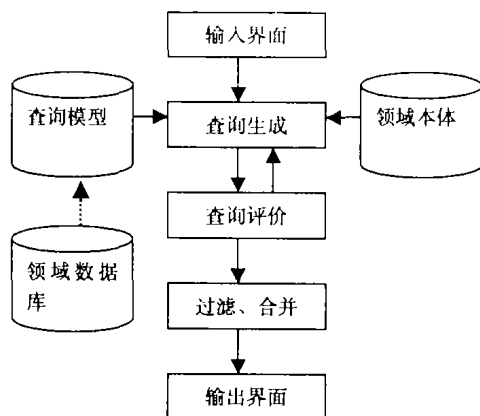


图1 基于本体的信息检索代理模板图

a. 输入界面。允许用户指明主题来执行检索,如一般的领域目录、出版日期。所有的这些数据组成一个咨询。我们视咨询为一个独立于数据库的摘要,概念化的高层次的查询。

b. 查询的生产和重组。咨询变成了查询生成模块的输入。查询生成模块是整个系统的核心,它跨越咨询、领域本体和查询模型,因此它能将一个咨询转变成低层次的、依赖于数据库的查询。在后

面会看到,在咨询与具体查询之间还有一个信息层,它建立在领域知识模型之上,我们称之为概念查询,它在咨询和具体查询之间起链接作用。另外,该模块在检索结果无效时,能重组具体查询条件。

c. 查询评价。对具体查询的检索结果——大致上是文献参考集合进行评价。该模块的功能是根据一些标准,给这些参考书目分配分值,作为它们实现用户规格或被文献引用的文献的事实质量的一个度量值。而且,对每一个概念查询,具体查询的结果在一个不同组内联合起来。

d. 过滤和合并。文献参考书目必须被过滤和合并以得到一个最终的一个分值。由于查询的特性——可能出现重复书目,这个过程是必须的。首先,要消除同一参考书目的不同出现形式;其次,建议删除那些不满足最小满意度约束标准的参考书目。

e. 输出界面。最后,检索结果经交互式的重新合并之后用一个合适的方式展示给用户。

f. 查询模型。指的是代表在不同抽象层次查询的信息方案。前面已经提到了两个层次,咨询一较高层次,具体查询一最低层次,这使我们能够让代理更加独立于语境。

g. 领域本体。包含了用来生成查询的一些领域知识。可用基于框架的表示方法,将它设计为一个层次树。该本体在一定程度上应该与语境无关,但应该指向用于查询生成模块的搜索引擎的元素。

2 专业数据库的构成

本系统中的专业数据库(或领域数据库)指的是关于某个领域知识的文档的集合,是以元数据为基础组织的。

元数据是关于数据的数据,即关于数据的内容、质量、状况和其他特性的信息。元数据是使数据发挥作用的重要条件之一,它帮助数据生产单位有效地管理和维护数据;帮助用户了解数据,以便就数据是否满足其需求作出正确判断;提供数据生产单位数据存贮、数据分类、数据内容、数据质量、数据交换网络及数据销售等方面的信息,便于用户查询检索。

与文本相关的元数据的一般形式有唯一标识符(UID)、作者、发表日期、发表源、长度、类型等。这种元数据常常被称作描述性元数据。另一种元数据刻画能够在文档内容中找到的主题内容,我们称之为语义元数据。

设专业搜索引擎允许通过称之为“检索限定词”(如 MAJR、TERM:NOEXP、TERM:TI、TW)的约束来实现对一个关键词执行不同类型的检索。不同的检索限定词对同一检索加以不同的约束,限制用于执行检索的数据字段集。如 TERM:NOEXP 表示可在领域本体的上位或下位词中进行检索,TI 表示在文章标题中进行检索,TW 表示所输入的值除了出现在标题,还出现在摘要中等等。

在后面会看到,我们的系统使用这些限定词对一个术语执行多个查询,通过领域本体来决定可用哪一类限定词。评价过程根据用在相关查询中的查询限定词给文章分配分值。

3 领域本体的构造

本体论是哲学上的一个概念,用来描述关于存在及其本质和规律。20世纪90年代本体论和本体的概念被引入计算机领域,人们利用本体来描述某个领域实体的存在本质。在人工智能领域,本体受到广泛的关注,已经和应用于知识组织与管理、信息资源规划、智能检索系统等方面。本体在信息系统中的应用,主要包括处理信息组织、信息检索和异构信息系统互操作问题。本体的目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并不同层次的形式化模式上给出词汇和词汇之间相互关系的明确定义。

本体的设计是一个创造性的过程,即使对于同一个领域,设计的本体都因人而异。目前没有一个标准的构造本体的方法,大家公认的一个规则是在构造特定领域的本体的过程中需要领域专家的参与。同时,设计本体要根据项目任务的需求进行,要求具有可扩展性和可维护性。本体设计是一个逐步完善的过程。下面介绍本体以及构建本体的基本过程。

首先说明一下本体的基本组成。一个本体包括一系列类(classes)或概念(concepts),它们是本体的核心,其定义一般采用框架结构,包括概念的名称、概念之间关系的集合,以及用自然语言对概念的描述。一个本体还包含用于描述有关概念的各种特征的属性(properties)和槽(slots),还包含槽的限制条件(restrictions)和分面(facets),以及一系列与某个类相关的实例(Instances),这些实例组成一个知识库(knowledge base)。

构建一个本体的基本过程有:定义本体中的类、定义槽并描述其允许的赋值,为实例的槽赋值。通过定义这些类的实例,建立起一个知识库。

3.1 定义类 无论采用哪种方法构建本体,都要从定义类开始。定义类可以采取“自上而下”、“从下而上”的方式。所谓“自上而下”,就是定义领域中的一些最抽象的概念作为类名,然后依次对抽象概念进行具体化。“从下而上”方式正好相反。当然也可以综合这两种方式。定义基本类后需要进一步完善类与类之间的分类等级体系。分类等级体系具有继承性,即子类继承上位类的全部槽和分面,因此同一个类中的若干直接子类要处于同一水平面上。同时,一个类可以是若干类的子类,要处理好类之间的多重继承关系。

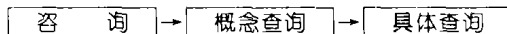
3.2 定义类的槽 通常类的定义不能提供足够的信息。在定义类的,还必须描绘概念之间的内存结构。例如,我们知道一个类可以是若干类的子类,同样,一个属性可能属于多个类。因此要将属性附于相应的类,于是类具有了自己特定的槽。通常,描述事物内部特征、外部特征的属性以及表示事物之间联系的属性都可以作为相应类的槽。由于任意类的所有子类均继承了该类的槽,因此一个槽应该附加在拥有该属性的最大的类上。槽可以有不同的分面来描述值的类型(value type)、允许的值(allowed values)以及值的基数(cardinality)。基数分面用来定义一个槽可以拥有的值的个数,类型分面则定义了该槽内可以进入何种类型的值。

3.3 创建实例 创建实例是本体设计的最后步骤。先选择要创建实例的类,然后针对该类填写槽值。

4 查询模型

为了理解下面章节,现在给出我们称之为查询模型的数据结构

的一些定义,因为它们描述的是关于查询的信息。首先,我们考虑到三层,正如在 Mesh 本体中一样。



第一层——查询,将所有需要的信息组合来执行查询。概念查询与本体中的领域类别直接相关,对包含在查询中的每一个领域类别来说都需要一个概念查询。

最后,提供具体查询来定义和捕获现实的、物理的查询结果。第一个结果都是将一些关键词和由本体提供的搜索术语合并、外加一些搜索标识符得到的结果。

这些模型促成了查询的创建和对它们的评价。为此,每一层包含该层下面的结构集合。它使得查询生成模块在查询生成过程中不断深入,而在评价和合并查询结果时停止作用。

这种将查询模型分解在三个抽象层次的思想有利于重用:用在其它搜索引擎和其它领域中。只有最低层依赖于具体的搜索引擎。高层完全独立,中间层具有依赖性仅仅因为它们与定义在本体中的领域学类别的关联。

4.1 咨询 咨询描述了用户的需要。用户在咨询窗口输入要查询的,包括一系列关键词、专业领域类别和一些其它的过滤条件,如起始年(将排除在这一年前的所有检索出来的文档)、终止年(排除在这一年之后的所有文档)等。

其中,关键词是搜索过程的核心,由表示主要的主题词或表达式构成。任何字符串都可以作关键词,但推荐使用有效的专业术语作关键词,从而使检索结果更有效。专业领域类别指的是前一节构造的领域本体的类别中选择一些类别,这就使得用户将查询集中于他们感兴趣的领域。所有这些类别的选择与其它类别可以是相互独立的。

4.2 概念查询 概念查询是一个介于咨询和具体查询之间的结构,它不依赖于搜索引擎。概念查询有如下元素:领域类别、具体查询(由查询生成器中的第二层分解得到,见第5节)和评分后的文档(在查询评价过程中生成)。

4.3 具体查询 具体查询中包括的元素有查询术语、查询概念(指向领域中的概念)、查询限定词、查询字段、检索出来的文档。这是一个低层结构,与搜索引擎共同工作,因此依赖于搜索引擎。

5 查询的生成和重组

下面描述将一个咨询转变为具体查询集的过程,以及这些查询条件是如何送往搜索引擎的。在详细解释该过程之前,先讨论一下生成查询任务的不同方法。

如前所述,咨询中隐含一些概念查询,对应于每一个领域类别。而且,我们知道,一个领域类别与一个专业术语和一些非专业术语相连。另外,每一个术语都可添加检索限定词。因此,系统不得不处理大量的信息。所以必须设计一个方法来使时间成本最小化,信息质量最大化。

可从三方面考虑来设计查询生成:a. 平行传送所有的查询条件,从而避免传送一个查询请求等待传送另外一个;b. 使用一种简短的检索格式,在生成和评价过程中仅仅检索文献索引(UIDS),从而节约时间和空间;c. 执行简短的查询条件,用几个检索术语而不是长的查询条件,这在进行 UIDS 时有用。因为评价过程能够据此出现在查询中对文档进行评分。

下面阐述一下在查询生成和重组过程中的分解过程(见图2)。我们视该过程为一个迭代的分解过程,从最抽象层——咨询到最低层次——具体查询层。后面我们会看到,评价和组合过程能倒序进

行,正如系统将不同查询合并成通用的对象一样。

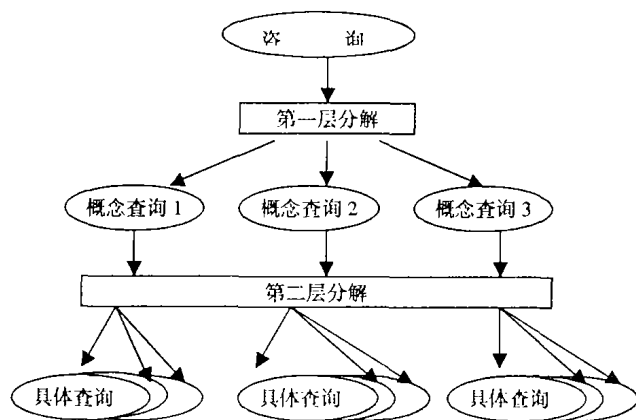


图2 查询分解图

第一步包括概念查询的生成,每一个对应用户选择的领域类别,另外其它的概念查询仅仅执行关键词和非领域类别的查询。每个概念查询有三个输入参数:一系列关键词,一个领域类别和具体的过滤条件。

第二步,查询生成器将每一个概念查询分解成一个个具体的查询。每一个具体的查询由关键词、表示领域类别项的概念查询的过滤条件及其允许的查询限定词。根据与概念查询相关的领域类别的结构,每个概念查询生成的具体查询的数目和类型差别很大。

关键词、具体过滤条件和与领域类别相关的术语由 AND 操作符,即限制性最强的一种操作符关联起来。评价结束后,如果结果不令人满意,生成过程将产生新的具体查询集合,但现在使用的是 OR 操作符。这时,我们仅使用了一个定量的标准,简而言之,如果检索出来的文档数目小于一个固定阈值,那么重新组织查询条件。

6 对查询结果的合并及评价

如前所述,对每个概念查询生成了一个具体查询集合。每个具体的查询与搜索引擎交互,检索出一系列文档。在本节,我们阐述如何将根据每个概念查询检索出来的文档进行合并,以及对文档进行评价的函数。

在概念查询内对文档评分可看成参照与之关联的领域类别,给每个文档分配一个成员值。因此,关键是理解定义在本体中查询概念的含义,以及运用于这些概念的限定词的意义。于是,我们区分用于成员意义的关系的次序:

查询概念(说明查询的类别):

如:领域术语 > 与领域相关的术语 > 与领域无关的术语 > 子标题

查询限定词(说明查询的子类型):

MAJR > TERM:NOEXP > TERM > TI > TW > no-modifier

为了给文档评分,我们为每个查询概念定义了一个数字型字段的数据结构和一个存储总分的字段。评分过程分两步:首先,在概念查询内为文档评分;其次,根据用户的选择组合不同概念查询的结果。

6.1 在概念查询内为文档评分 检索完毕代理得到在不同具体查询中检索到的多个系列文章(仅 UID)。现在代理需要将它们以某个次序合成一个系列。为实现该目的,系统使用了一个聚合函数。

6.1.1 聚合函数需要计算文档 a 在概念查询 j 中的重要程度,该重要程度依赖于其在不同具体查询结果中出现的次数。聚合函数定义如下:

对一个概念查询 j : $\Theta_j(a) = \sum_i c_i \theta_{ij}(a)$

其中, a 是检索到的文档, C_i 是每个查询限定词 i 的权重系数。权重系数满足如下限制条件:

$$\sum_i c_i = 1; \quad \forall_i: c_i \geq 0$$

θ_{ij} 是在概念查询 j 中查询限定词 i 得到的文档的成员函数。

6.1.2 成员函数

$$\theta_{ij}(a) = \frac{\sum_k b_k n_{ijk}(a)}{N_{ij}}$$

其中, $n_{ijk}(a)$ 是在概念查询 j 中,使用限定词 i 和查询概念 k 文档 a 在具体查询中出现的次数。 N_{ij} 是一个规范化的系数,其定义如下:

$$N_{ij} = \sum_k b_k N_{ijk}$$

N_{ijk} 是每个概念查询 j 使用查询限定词 i 和查询概念 k 的具体查询的数目。它代表了 n_{ijk} 的理论上的最大值。对每个查询 k 权重系数 b_k , 必须满足条件: $\forall_k: 0 \leq b_k \leq 1$

权重系数值可根据经验估计,设置为固定的值,理想状态下可让代理通过与用户交互来学习这些权重系数。

6.2 合并来自不同概念查询的文档

此时,对每一个概念查询,系统已有一些经过评分的文档。允许用户组合来自不同概念查询的结果。为此,我们要比较和组合这些文章的分值。

这并不是一件容易的事,因为概念查询与领域类别相关,而领域类别结构不同。其结果依赖于用于领域类别的术语的个数,这些术语的特征和类型。其结果是并非所有的文档都容易得到较高的分值,而是依赖于此概念查询被检索的地点。对某个概念查询最好的文档分值为 0.8 与最好的文档分值为 0.2 是不一样的。

我们想用一個函数根据经验中的最大值来修正对文档的评分。对获得最大分值的类别无须修正,因此这些类别的最大值较低,我们使用一个较大的修正值。为实现该目标,我们定义了如下函数:

$$\Theta(a) = \sum_j \frac{\Theta'_j(a)}{N} \quad \Theta'_j(a) = (\Theta_j(a))^{K_j}$$

其中, N 是我们欲组合的概念查询的个数, K_j 是 $\Theta_j(a)$ 的最大值。用新得到的评分值,对将要显示给用户的结果文档进行排序。

一般的信息检索函数是基于词汇的频率来比较文档和查询条件,这对非结构化的数据最适宜,当然也能运用于结构化的数据,如本系统中提到的专业领域数据库。问题在于系统不得不检索、存储和分析大量的信息,要消耗大量的资源,所以并不实用。本系统则避免了该问题:不是一次检索全部文档的内容来给它评分,而是对每个概念生成大量不同的查询,仅检索数据库中同样数目的文档。评分过程基于文档出现的概念的数量和特征。因此作者认为该方案充分利用了 Internet 的性质,因为可修改为并行执行检索过程。

当然文中仅仅提出了一个设计方案,下一步的工作是结合某个领域进行具体实现。

参考文献

- 1 J. Brutlag, J. Meek. Challenges of the Email Domain for Text Classification. in Proceedings of the 17th International Conference on Machine Learning ICML00. 103 - 110, 2000
- 2 Jose Maria Abasolo, Mario Gómez MELISA. An Ontology-based Agent for Information retrieval in Medicine. <http://www.ics.forth.gr/proj/isst/SemWeb/.proceedings/session3-1/paper.pdf>.
- 3 万捷等.本体论在基于内容信息检索中的应用.计算机工程,2003;(4)
- 4 邓志鸿等.ontology 研究综述.北京大学学报(自然科学版),2002;(38)

(责编:勤王京)