中 文 信 息 学 报
第19卷 第2期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 19 No. 2

文章编号: 1003- 0077(2005) 02- 0051- 07

基于 Ontology 的信息检索技术研究①

陈 康, 武港山

(南京大学 计算机软件新技术国家重点实验室, 计算机科学与技术系, 江苏 南京 210093)

摘要: 随着 Web 的迅速发展, 网上信息资源越来越丰富, 网络已经成为了一个全球最大的信息库。而用户要从中得到所需的信息一般是通过各种信息检索工具。但是现有的信息检索工具都存在着检索精度不高等问题。本文针对这些问题, 提出了将 Ontology 融合到信息检索技术中的思路。利用 Ontology 中拥有的领域知识, 可以大大提高检索系统对自然语言文本的理解能力, 同时方便用户以自然语言的方式提出检索请求, 从而提高检索的效果。

关键词: 人工智能; 自然语言处理; 信息检索; Ontology; 自然语言理解

中图分类号: TP391 文献标识码: A

Research of Ontology-based Information Retrieval

CHEN Kang, WU Gang-shan

(State Key Laboratory for Novel Software Technology,
Departmentof Computer Science & Technology of Nanjing University, Nanjing, Jiangsu 210093, China)

Abstract: Along with the rapid development of Web, the information resources in the web are becoming more and more abundant. People get information from Web mainly by search tools, but always puzzled by the precision of them. To solving this problem, we adopted domain Ontology in our information retrieval system. By using of the domain knowledge in Ontology, retrieval system could improve semantic understanding of retrieved documents, and give the chance to user to put their information request in more nature way (more precise way). Experimental results show this method can increase the precision of information retrieval.

key words: artificial intelligence; natural language processing; information retrieval; Ontology; natural language understanding

1 引言

随着 Internet 的快速发展, Web 已经发展成为全球的信息源。用户从这个信息源中获取信息一般都是通过搜索引擎来进行的。但是使用传统的搜索引擎, 用户要精确地找到所需要的信息往往十分困难, 这主要有几个方面的原因。第一是对用户的问题理解不够准确, 导致返回结果中含有很多噪声, 用户不能够很容易的找到自己所需要的信息; 第二是对信息内容的处理大多采用的是基于某种编码过程的预处理技术或某种全文分析技术, 仅仅反映内容的一个侧面; 第三是用户提出的问题与信息源的内容不可能完全一致, 难以保证内容与用户问题正确匹

① 收稿日期: 2004- 06- 20
基金资助: 国家自然科学基金资助项目(6007303); 国家"863"计划资助项目 (2002AA117010- 10)
作者简介: 陈康(1981—), 男, 硕士研究生, 研究方向为信息检索.

配,正确率很低。要提高现有检索系统的精度,就必须解决好上面提到的这几个问题。现有的一些研究工作表明[1~4],基于 Ontology 的技术是解决这些问题的方法之一。

文献[1]通过建立一个基于 WordNet[5]的 Ontology,解决了从黄页和产品目录中进行信息检索的问题,但是黄页与产品目录信息一般来说都具有一定的结构,因此在文献[1]中所采用的技术不能够很好的应用于自然语言文本的检索;文献[2~4]主要研究通用 Ontology 在信息检索及自然语言理解中的应用,但是要建立一个能够涵盖所有领域知识的通用 Ontology 是很困难甚至是不可能的,因此比较现实的方法是建立某个领域的 Ontology,利用它去解决该领域的特定的信息检索问题。

Web 是一个开放的信息空间,信息内容涵盖社会生活的方方面面,表达方式也主要是非结构化的形态。因此,基于开放领域的 Ontology 来解决 Web 的信息检索问题,还有许多实际难以操作的地方,如:Ontology 的建立、开放领域的自然语言理解问题等。本文的研究思路是利用特定领域的 Ontology,采用传统的分类等技术,过滤收集 Web 中该领域的信息内容,允许用户以自然语言的方式表达自己的检索请求,提供该领域内的高精度信息检索服务。

在本文中,我们以奥运为背景建立了一个奥运领域的 Ontology,并着重讨论了基于 Ontology 的自然语言检索关键技术的处理方案。

2 基于 Ontology 的信息检索处理技术

传统的信息检索方法主要分为两大类,第一类是基于关键词匹配的方法,这种方法首先让用户以关键词的形式提出检索请求,然后将用户提交的关键词与文档库中的文档进行匹配,最后将那些出现了用户所提交的关键词的文档作为检索结果返回给用户。现在 Web 上很多搜索引擎都是采用这种检索方法,例如 Google[6]、百度[7]等。但是这种方法最大的一个不足就是其检索过程中不包含任何语义信息,这是导致检索精度不高的一个很重要的原因。第二类方法称为概念信息检索[8],它通过对文档中的信息进行语义层次上的处理来析取各种概念信息,并由此形成一个概念库,然后根据对用户的问题的理解来检索概念库中相关的信息以提供检索的结果。这种方法克服了基于关键词检索中不考虑语义信息的局限性,并且具有较好的自然语言接口。但概念信息检索一个不足之处就是其概念库中不包含概念间关系的描述,因此无法处理有关概念关系的问题。

Ontology 是对概念化的明确描述,它把现实世界中的某个应用领域抽象成一组概念及概念间的关系[9]。我们把 Ontology 融合到传统信息检索技术中去,不仅可以继承概念信息检索的优点,还可以克服概念信息检索不能对概念关系进行处理的局限性。

在本文中,一方面我们利用领域 Ontology 对该领域内的自然语言文本进行分析,并利用分析结果对 Ontology 中的概念进行实例化——我们将这些实例化了的概念称为信息实体,然后根据 Ontology 中概念之间的关系将这些信息实体组织起来,这样就将原来的半结构或无结构的自然语言文本转化成了具有一定结构的信息实体及这些实体之间的关系。另一方面,对于用户提出的检索请求,我们也利用领域 Ontology 对其进行理解,将其转化为对某个信息实体及其属性的查询,保证了用户问题与信息描述的一致性,可以实现他们的精确匹配,从而达到提高检索精度的目的。

下面将针对如何在信息检索中应用 Ontology 技术进行详细阐述。

3 基于 Ontology 的检索请求处理

传统的信息检索工具提供给用户的主要是基于关键字的检索接口,但是在很多情况下用户真正的检索意图很难用几个关键字表达清楚,这也是导致现有检索系统的精度不高的原因之一。因此为了能够更好的让用户表达出他的检索意图,我们提供给用户的检索接口是自然语言的表达方式。用户可以以自然语言的方式向系统提出问题,例如某个用户希望知道姚明的身高,他可以提出问题"姚明的身高是多少?",此时我们利用领域 Ontology 中的知识和一些简单的自然语言理解的技术对用户的问题进行分析,得到用户真正的检索意图,然后将检索请求提交给系统的检索部分。问题处理过程如图 1 所示。

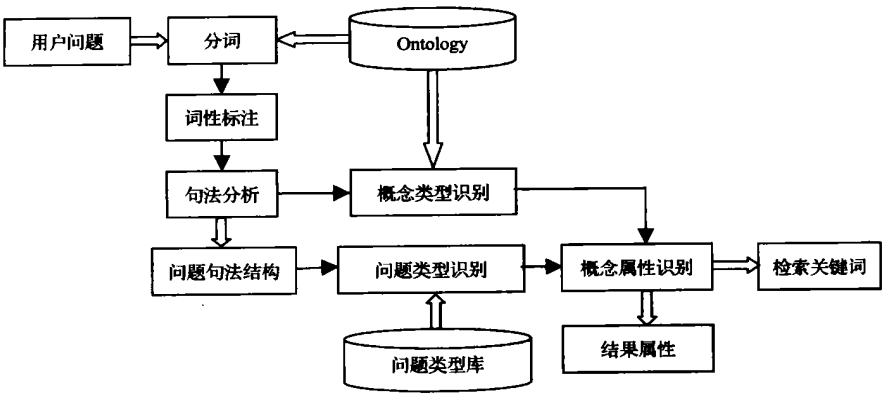

图 1 问题处理框架图

对于问题的处理,首要问题就是分词。在进行分词的时候我们使用抽词工具并结合领域 Ontology 中包含的领域专业词典识别领域专有人名、机构名、专业术语。对于分词结果还需要进行词性标注,以便在此基础上进行问题的句法分析。

概念类型识别的作用是根据句法分析的结果和领域 Ontology 中的概念类型模板,识别出该问题所描述的概念的类型。例如:对于问题"姚明的身高是多少?",通过概念类型识别之后我们可以知道该问题所关心的是人物这个概念中的某个属性,这样我们在进行检索的时候就可以只分析那些属于人物概念的信息实体,从而减少信息检索的处理时间。

问题类型的识别是指将用户的问题根据问题类型库划分到一个指定的类型中[10,11]。在进行检索之前,识别该问题的类型可以简化检索的工作,同时使得检索结果更加趋于精确。例如问句"2004 年奥运会在哪里举行?"是要寻找"地点"类型的答案,那么这个问题的语义类型就是有关地点的,在检索时只需要寻找"地点"类型的属性信息。识别问题类型的方法有很多,例如基于句法分析、基于机器学习的分类算法等。这里我们采用基于问句句法分析和语义分析相结合的识别方法:首先,为每个语义类型建立若干条对应的规则模板;然后将问题进行句法分析后的结果和规则模板进行匹配,识别问题的类型。如果有多个模板与问题相匹配,则采用最大匹配原则进行筛选。对于通过句法规则模板匹配不能判定语义类型的问题,需要利用领域 Ontology 中包含的词汇语义知识进行辨析。例如问题"射击比赛在什么地方举行"与问题"射击比赛在什么时间举行?"两个问题句法结构均与模板"什么/r + NP"相匹配,但是显而易见两个所要表述的语义内容全然不同,问题的类型取决于"什么"后紧跟的名词短语的语义内

容,这时我们就需要结合领域 Ontology 中所表述的词汇的语义知识,分析判断问题的类型。

得到问题的概念类别和类型之后,通过概念属性识别我们就可以知道用户需要查询的是什么概念中的哪个属性。例如对于问题"姚明的身高是多少?",经过上面的处理之后我们就可以知道用户查询的是人物概念中的身高属性,那么检索结果就必须是人物实体中的身高属性。

最后我们要从用户问题中提取出检索关键词并将它们提交给系统的检索部分。但并不是问题中的所有词语都可以作为信息检索的关键词,首先我们将问题中出现的一些没有实际意义但是使用频率很高的虚词和功能词(如"的","了","把"等)过滤掉,然后根据领域 Ontology 中的知识提取出问题中包含的领域主题概念,找出问题中与主题概念表述相关的词语,最后将这些词根据同义词典进行适当的扩充后得到检索关键词提交给系统的检索部分。例如从问题"姚明的身高是多少?"中我们可以提取出关键词"姚明+ 身高| 高度"。

4 基于 Ontology 的文本预处理

信息检索的目的,就是根据用户的检索要求,从大量的信息中找到满足用户要求的信息,并对检索结果按照与用户请求的相关性大小进行排序后返回给用户。要从大量的信息中查找所需的信息,如果不对文本进行任何处理,仅仅通过字符串匹配这种方法,效率肯定十分低下,因此必须对文本信息进行一定的预处理,以达到快速和准确的检索的目的。

本系统中,文本预处理的目的是从非结构化的文本信息中提取出文本中的有用信息并根据领域 Ontology 的概念类型模板形成信息实体,从而将这些非结构化的文本信息转化成具有一定结构的信息实体。文本预处理过程如图 2 所示。

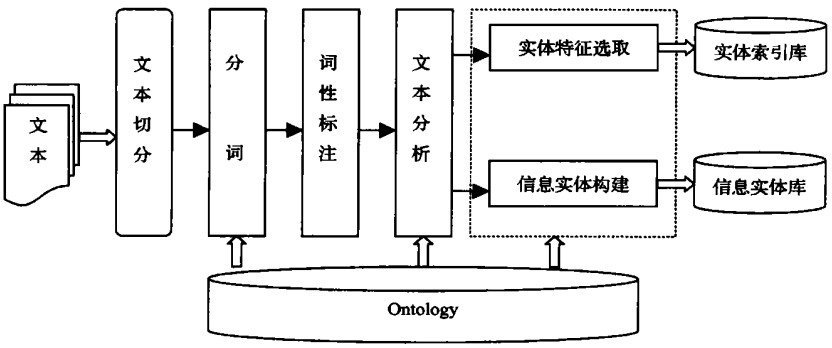

图 2 文本预处理流程框架图

一般的信息检索返回的是与用户查询相关的整个文本的信息,但在这些文本中,可能只有其中一小部分才是所需要的结果,检索时只需要返回和检索请求最为相关的小段文本。同时在利用领域 Ontology 构建信息实体时,因为大段文本包含的信息量比较多,构建信息实体时处理过程比较复杂。因此,在对文本内容进行分析处理之前,我们事先将整篇文本划分若干小段文本,然后再进行分词与词性标注的处理。

根据领域 Ontology 中的知识构建出大量信息实体的主要目的是为了提高后续信息检索的匹配精度和检索速度。例如如果我们能够事先从文本信息中构建出"姚明"这个人物实体,并根据领域 Ontology 中的领域抽象概念类的描述特征信息,确定这个人物概念的身高、体重等各种属性信息,那么当用户提出问题"姚明的身高是多少?"的时候,我们可以直接从这个信息实体中找到检索结果返回给用户。

对于信息实体的构建我们采用的是一种基于自然语言理解的信息抽取技术。首先,我们结合领域 Ontology 中包含的领域专业词典对文本进行词法分析,识别领域专有人名、机构名、专业术语,然后在此基础上进行句法分析,找出该段文本所描述的对象,并根据领域 Ontology 中的领域抽象概念类的描述特征信息,确定出该文本内容的领域概念类型,即确定该实体的概念类型,例如对于句子"姚明的身高是 2.26 米。",通过处理,我们可以确定该句子描述的是人物这个概念;对于信息实体中属性信息,我们采用的是基于规则的信息抽取方法,首先我们定义了许多规则,例如对于人物的身高属性我们定义了规则"{身高/n 是/v (/m [/q] [/m])|身高/n (/m [/q] [/m])|高/a (/m[/q] [/m])|身高/n 为/v (/m [/q] [/m])}",我们将文本经过词法分析之后的结果与这些规则进行匹配,就可以得到文本中包含的信息实体的属性信息。例如上面的句子经过词法分析的结果是"姚明/nr 的/u 身高/n 是/v 2.26/m 米/q",然后我们根据身高属性的规则就可以得到人物"姚明"的身高属性的是"2.26 米"。信息实体构建完成后将其存放到实体库中,便于将来进行检索时使用。

5 信息实体的索引与检索

信息检索的最终目的是让用户能够快速而准确的得到其所需的信息,而系统的信息索引模型和检索机制的好坏则会直接影响整个系统的性能,因此对于信息检索系统来说,一个好的信息索引模型和检索机制是必不可少的。

5.1 信息实体的索引

对信息实体进行索引的首要工作就是要进行信息实体特征项的选取。实体特征项可以是文本中的各种语言单位,对于中文来说可以是字、词、短语,甚至是句子或者句群等更高层次的单位。因此,特征项的选择只能由处理速度、精度、存储空间等方面的具体要求来决定。选出的特征项越具有代表性,语义层次越高,所包含的信息就越丰富,但是分析的代价就越大,而且受分析精度(如句法分析的正确率)的影响就越大。由于词汇是文本最基本的表示项,在文本中出现的频度较高,且呈现出一定的统计规律,同时考虑到处理大规模真实文本所面临的困难,选择词或短语作为特征项比较合理,常被应用于文本检索领域,本文中也采用这种方法。但是并不是文本中所有的词都可以作为实体特征项,文本中存在一些没有实际意义但是使用频率很高的虚词和功能词(如"的","了","把"等),常常把一些真正具有表征意义的实词淹没掉,因此我们将虚词、语气词、介词、连词、特高频率词、特低频率词等组织成一个停用词表(stop list),把表中的词汇从特征集中滤掉。

在实际的检索过程中,为了能够通过特征项快速查找到信息实体,我们设计了如图 3 所示的数据结构。文本信息中所有的特征项用一个链表连接起来,每一个特征项节点指向一个信息实体的链表,实体链表中的每个节点记录实体库中的实体号。通过实体号我们就可以很方便的从实体库中查找到该实体的信息。

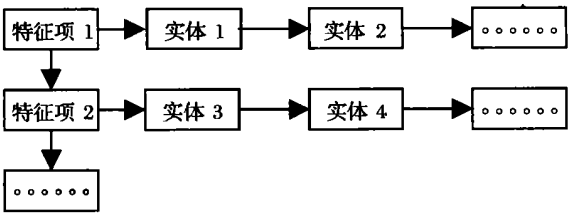

图 3 实体索引结构

5.2 信息实体的检索

在文本预处理阶段,我们将自然语言文本根据领域 Ontology 中的知识转化成大量的信息实体;在问题处理阶段,我们将用户的问题转换成对某个信息实体的属性的查询,通过这两部

分的处理之后我们就将自然语言检索的问题转换成了对信息实体检索的问题。

信息实体的检索分为三个阶段。首先,检索部分得到从问题处理部分提交过来的检索关键词,把关键词都作为信息实体的特征项,寻找它在特征项链表中的位置。如果找到了,那么沿着该特征项所指向的实体链表,可以统计出哪些实体包含该特征项。对每一个关键词都进行这样的处理后,会得到多个实体号;然后,我们去掉重复实体号之后,到信息实体库中就可以查找到实体的所有信息。但是本文中,实体特征项是用词语来表示的,而词语的多义及分词过程中的歧义现象会导致同一个特征项返回多个信息实体,因此我们必须对返回的信息实体进行筛选。因为在问题处理部分,通过概念类型识别我们能够得到用户问题所关心的实体的类别,所以在进行实体筛选的时候,我们把那些与用户问题的概念类别不相同的信息实体过滤掉;最后,把剩下的信息实体按照它们所包含的检索关键词的数量进行排序,根据在问题分析阶段中的分析出来的结果属性,提取出实体中对应属性的值作为检索结果返回给用户。

6 基于 Ontology 的 Web 信息检索系统

我们将本文的思想应用于 Web 信息检索中,实现了一个基于 Ontology 的 Web 信息检索系统,该系统包含五个主要处理模块:Ontology 管理模块、问题处理模块、文本预处理模块、信息检索模块、库文件管理模块,其系统结构如图 4 所示。

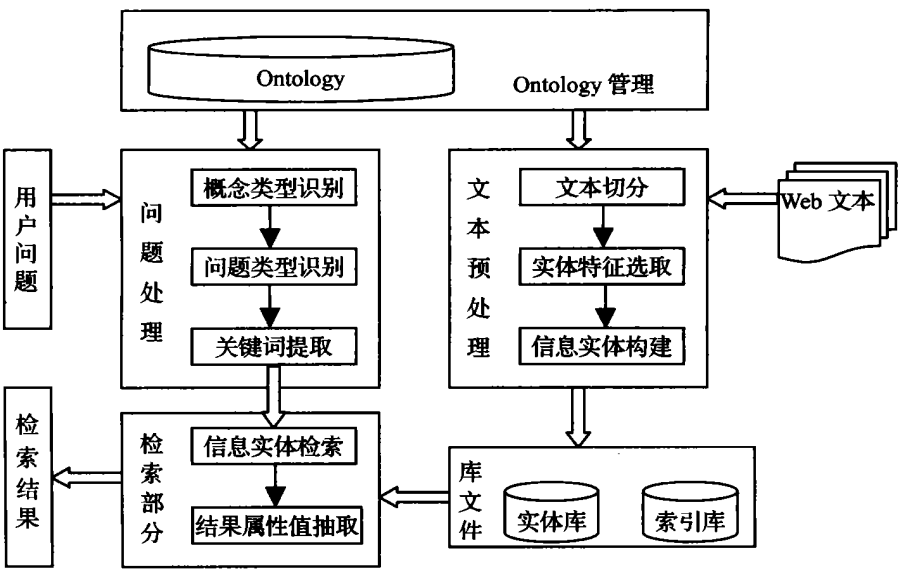

图 4 基于 Onrology 的 Wed 信息检索系统框架

图中的各个模块相互协作,共同完成用户问题的回答任务。对于用户来讲,他直接把问题提交给问题处理模块,然后等待处理的结果。系统在接受了用户的问题后,首先进行问题分析和处理,得到检索关键词和问题类型。检索关键词将提交给检索部分去获取信息实体,问题类型将被转换为结果属性提交给检索模块,用于从信息实体中提取结果属性值。检索模块提取出实体的结果属性值以一定的形式返回给用户。

我们用本系统和一个基于关键词的检索系统分别对我国 32 位优秀冰雪运动员的简历进行了处理,并做了一个简单的性能比较:例如,某个用户希望知道运动员"罗致焕"的年龄,向基于关键词的检索系统输入关键词"罗致焕+ 年龄",我们发现没有文档返回;而向本系统输入问

题“罗致焕的年龄是多少?”,系统将罗致焕的出生年月读取出来,并同时把包含这个信息的文档返回给用户;而如果用户向基于关键词的检索系统仅输入关键词“罗致焕”,我们发现返回了两篇文档,其中一篇与本系统返回的相同,而另一篇文档并不包含用户所需要的信息。从这个简单的性能比较我们可以看出本系统确实在一定程度上提高传统检索系统的检索精度。

7 结束语

Ontology 是对世界或者领域的概念化描述。本文把 Ontology 应用到传统信息检索技术中,主要从两个方面提高了检索系统的能力。第一是在对用户问题进行理解的过程中,利用领域 Ontology 的知识,将用户问题转换为对领域 Ontology 中某个概念的相应属性的查询,缩小了信息检索的范围,从而减少信息检索的处理时间;第二是在对文本进行分析的时候,利用领域 Ontology 进行文本特征项的选取,同时根据领域 Ontology 中概念类型的描述模板完成信息实体的构建,在进行检索时可以直接从这些信息实体中查找满足用户需求的信息,从而提高检索的精度。

我们通过所实现的原型系统,发现还有以下几个问题需要解决:首先,如何建立 Ontology 还没有一个很好的理论支持,并且 Ontology 中的概念一般都是通过人工提取的,这使得基于 Ontology 的应用不能大规模开展,因此需要开发出能够自动或半自动提取概念的工具;其次,利用领域 Ontology 进行实体构建时,需要从文本中抽取出实体的属性,而本文中采用的基于规则和模板的信息抽取的方法并不能够很好的胜任这一工作,因此需要开发出基于语义的信息抽取工具。对这些问题都有待进一步的研究。

参 考 文 献:

[1] N. Guarino, C. Masolo, and G. Vetere. OntoSeek: Content-Based Access to the Web[J]. IEEE Intelligent System, 1999, 14(3): 70-80.

[2] 武成岗,焦文品,田启家,等. 基于本体论和多主体的信息检索服务器[J]. 计算机研究与发展,2001,38(6):641-647.

[3] 潘宇斌,陈跃新. 基于 Ontology 的自然语言理解[J]. 计算机技术与自动化,2003,22(4):71-74.

[4] 廖乐键,曹元大,李新颖. 基于 Ontology 的信息抽取[J]. 计算机工程与应用,2002,23(4):110-113.

[5] G. A. Miller. WORDNET: A Lexical Database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[6] Google[Z]. http://www.google.com.

[7] 百度[Z]. http://www.baidu.com.

[8] 何绍义. 概念信息检索的理论与实践[J]. 情报学报,1995,14(2):134-141.

[9] Gruber T R. A Translation approach portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[10] 郑实福,刘挺,秦兵,等. 自动问答综述[J]. 中文信息学报,2002,16(6):46-52.

[11] Seung-Hoon Na, In-Su Kang, Sang-Yool Lee. Question Answering Using a WordNet-based Answer Type Taxonomy [A]. Proceedings of the 11th Text Retrieval Conference (TREC-11)[C].

[12] Rohini Srihari, Wei Li. Information Extraction Supported Question Answering[A]. Proceedings of the 8th Text Retrieval Conference(TREC-8)[C].