《智能信息处理》课程考试

基于本体的信息检索模型研究

李怀清

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 18 日

基于本体的信息检索模型研究

李怀清

（大连海事大学 信息科学技术学院，大连 116026）

摘　要 文章首先对本体进行了简要的概括根据文档处理方式的不同，将基于本体的信息检索系统分为基于知识库的语义检索系统和基于语义网文档的信息检索系统两类。对这两个模型的实现原理和关键步骤进行了阐述;探讨并指出当前研究中存在的不足;最后对基于本体的信息检索系统的研究热点和方向进行了展望。

关键词:本体;信息检索;语义标注;知识库;语义网文档

Research on ontology-based information retrieval system models

Li Huaiqing

Abstract: The article first briefly summarizes the ontology. According to the different document processing methods, the ontology-based information retrieval system is divided into two types: the semantic retrieval system based on the knowledge base and the information retrieval system based on the semantic web document. The realization principles and key steps of these two models are expounded; the deficiencies in current research are discussed and pointed out; finally, the research hotspots and directions of ontology-based information retrieval systems are prospected.

Key words: ontology; information retrieval; semantic annotation; knowledge base; semantic Web documents(SWD)

0 引言

本体作为一种能在语义和知识层次上描述信息系统的概念模型建模工具,具有良好的概念层次结构和对逻辑推理的支持[1]。它在计算机领域中的应用使信息检索从基于关键词的层面提高到基于知识(或概念)层面上成为了可能。将本体融合到传统信息检索技术中,不仅可以对文档中的信息进行语义层次上的处理,还可以结合用户的检索条件利用 Web 上的语义信息进行推理,进而得到较为准确的结果。

1 基于本体的信息检索系统的分类

近年来,美国、欧盟等语义网研究机构和大学实验室相继设计和提出了不少有代表性的基于本体的信息检索系统,如基于语义网检索的 Metalog;最早基于顶层本体设计的 WebKB;基于 XML 表示的 Quest、Elixir、XIRQL 等。

这些系统开发的设计理念和侧重点不尽相同,没有明确的分类方法对这些系统进行界定。虽然不少信息检索系统引入了本体的概念,但是不同的信息处理方式导致研究学者在论述基于本体的信息检索系统时,经常混淆本体在系统中扮演的角色。

本文根据对文档处理方式的不同,将基于本体的信息检索系统分为基于知识库的语义检索和基于语义网文档的信息检索两类[2]。基于知识库的语义检索系统主要利用自然语言处理技术。根据领域本体描述将网页或自然语言文本转换为大量信息实体。这种信息实体以某种知识表示语言描述存储在知识库中,搜索引擎可以对知识库进行推理和检索。早期的 SHOE 项目、欧洲科研信息系统 AURIS-MM 以及 OntoText 语义研究实验室开发的 KIM 平台等都是基于知识库的语义检索系统的代表。基于语义网文档的信息检索系统的处理对象主要包含语义标注语言的网页,由语义网语言书写的语义网文档能被软件代理直接访问。它将语义网文档中的语义信息转换为搜索引擎能够处理的统一格式,存储在一个 RDF 文件或 OWL 文件中。这类系统包括 Ontobroker、马里兰大学设计和研发的基于语义网搜索引擎原型系统 Swoogle 以及 UMBC 大学 eBiquity 实验室开发的语义网信息检索、推理引擎 OWLIR 等。从两类系统的划分依据上可以看出,基于知识库的语义检索系统采取了向前兼容的策略。所谓向前兼容是指尽可能维持现有 Web 内容的形式,利用知识表示技术建立庞大的知识库,在已成熟的互联网搜索技术上进行有益的改进。基于语义网文档的信息检索系统采取向后兼容的策略,即其实验平台是 BL 等语义网学者推崇的语义网,代表着互联网的发展方向。

2 基于本体的信息检索模型

2.1 基于知识库的语义检索模型

基于知识库的语义检索模型(图 1)首先建立基于领域知识的本体库对文档进行预处理,建立本体库中实例与文档的链接关系。根据用户提交的请求检索知识库,对实例中的隐含信息进行推理,返回符合查询条件的文档集合。检索的结果经过排序处理后返回给用户。

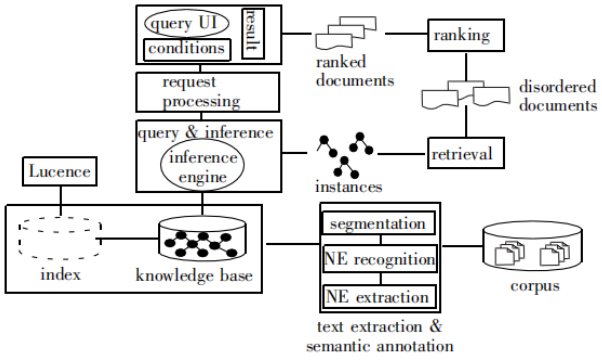

图 1 基于知识库的语义检索模型

2.1.1 构建领域本体库

本体的目标是捕捉相关领域的知识,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,并从不同层次的形式化模式上给出词汇和词汇间相互关系的明确定义。通常需要在领域专家的帮助下建立基于领域概念知识的领域本体[3]。

由于本体工程到目前为止仍处于相对不成熟的阶段,每个应用都拥有自己独立的方法,比如基因学专家可以根据本领域的专业知识建立对基因学的概念描述。有一些科研 机构正致力于领域本体标准的制定工作,通过标准的制定和实行,促进本体定义的规范及加强本体的可重用性。目前,比较有影响的本体标准包括 Dublin core、FOAF、SKOS core、CERIF 等。领域本体库的建立包括人工和自动的方式,它为文本抽取和语义标注以及查询请求处理提供参照知识,方便对知识进行格式转换和存储。

2.1.2 文本抽取和语义标注

文本抽取和语义标注的目的是从非结构化文本信息中提取出文本中有用的信息,并根据领域本体的概念类型形成具有一定结构的信息实体[4]。在对文本内容进行分析处理之前,事先将整篇文本划分成若干小段文本;然后进行分词与词性标注的处理,并且在分词过程中进行概念的实体描述和逻辑关系的提取。在检索过程中,查询接口返回的结果是本体库中的元组,而用户希望得到的是包含关键字的文档。所以,文本抽取和语义标注模块的另一个功能就是建立本体库中元组实例与文档的映射关系。每个实例包含一个标签属性,标签的值描述了实例的同义信息。通过启发式算法将文档中的实体与知识库中的实例进行匹配。通常使用文档—实例关联表来存储文档和实例间的对应关系,有了关联表,通过查询接口返回的元组实例就可获得相应的文档链接了。

2.1.3 查询请求处理

为了更好地让用户表达出他的检索意图,查询接口负责将用户提交的自然语言查询语句转换为合适的本体查询语句。

用户以自然语言的方式向系统提出问题;然后利用 ontology 领域中的知识和一些简单的自然语言理解技术对用户的问题进行分析,提取主题词,得到用户真正的检索意图;最后将检索请求提交给系统的检索部分。

在进行处理的过程中,首要问题就是建立本体库,然后对用户的问题进行概念类型识别和问题类型识别。概念类型识别的作用是根据句法分析的结果和领域本体中的概念类型模板识别出该问题所描述的概念类型。概念类型识别之后可以知道该问题所关心的是某个概念中的某个类或者属性。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。在用户提交问题后,系统就需要结合领域本体中所表述的词汇的语义知识分析判断问题的类型;得到问题概念类别和类型之后,系统就可以根据主题词库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。

2.1.4 索引与检索

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[5]。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索引文档类型、处理效率、存储空间等方面的具体要求来决定。

检索时,推理模块能够对本体库中用 RDF、RDFS、OWL 等语言书写的实例进行推理。推理过程还可以根据一定推理规则进行,系统管理员可以根据具体需要创建适合的推理规则。当检索系统返回元组后,通过查找文档—实例关联表便可以得到文档列表。对文档列表进行排序选择,最终返回给用户关联度较高的文档结果集。

2.2 基于语义网文档的信息检索模型

基于语义网文档的信息检索模型(图 2) 与目前流行的 Web 搜索引擎模型非常相似,其不同之处在于:

a) 该模型抓取的网页主要是带有语义标记的语义网文档而不是通常所说的 HTML 网页;

b) 索引类型不仅包括单词、词组、N-gram 等传统索引类型,还包括 SWD 的元数据类型,如三元组节点、URI 链接等。

复合型索引方法使传统的检索技术和基于本体的推理技术融合成为可能。

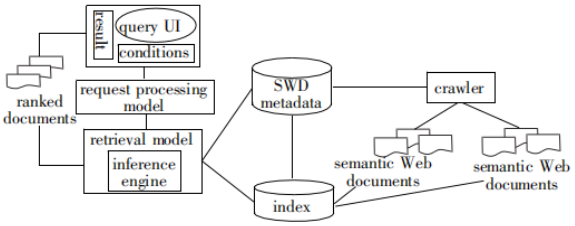

图 2 基于语义网的信息检索模型

2.2.1 抓取语义网文档

同传统搜索引擎的爬虫程序抓取 Web 中的网页一样,该模 型也需抓取 Web 中的 SWD。但目前大多数网站上的网页仍然是以 HTML 为主, 只有个别科研或者语义网实验平台上的网页属于 SWD。所以爬虫程序只需抓取后缀名为.rdf、.owl、.daml、.n3 的网页。需要指出的是,并非所有为.rdf、 .owl 的文档都是 SWD。有人对语义网文档在语义标记文档中所占的比重进行统计。结果表明以 rdf 为扩展名的语义标记文档中, SWD 占总数的 60%;以 owl 为扩展名的语义标记文档中,SWD 占总数的 67%。所以爬

虫程序需要对抓取到的语义标记文档进行 SWD 类型验证。

Web 搜索引擎利用网页间的 URL 抓取分布在 Web 上的网页,语义网文档爬虫则通过分析 SWD 间的语义关系来遍历 Web。SWD 中通常包含大量的 URI,这些 URI 隐含的命名空间通常指向另一篇 SWD 的 URL;OWL 的 import 关键字说明其导入的本体所属的文档也是一篇 SWD。

2.2.2 索引和检索

目前,Web 上的语义网文档通常由纯文本与语义标记混合构成。所以,传统的基于关键字的索引技术仍可以应用在基于语义网文档的信息检索模型中。除了对单词、短语、句子等类型建立索引外,语义标记特征项或者 URI 也可以成为索引的对象。索引建立好后, 搜索引擎便可以进行检索了。在检索过程中, 运用本体的推理机制,具体过程与基于知识 库语义检索模型的推理功能相似,不过后者提供完整的知识库。基于语义网文档的信息检索模型通常直接对语义网文档中的语义标记进行推理或者从文本文档中抽取出标记三元组存储到一个 RDF 或 OWL 文件中,对文件进行推理。

3 基于本体的信息检索研究的不足

1) 本体评价缺乏统一的标准。前面已经简单介绍了本体的一些构造准则, 但是这些评价准则基本是类似定性的描述,还没有定量、明确可操作的定量评价准则。如果不能解决好本体评价的问题,未来语义网中的本体定义标准繁多,对同一个概念的描述存在不同版本,这无疑违背了本体论倡导的知识共享的初衷。

2) 现有系统对新知识的更新支持不够。网络环境下,用户的信息需求很宽泛,特别是时代感很强,关注的内容与社会新闻和事件常常紧密相关。在基于知识库的信息检索系统中,本体库在领域专家的帮助下通过手工或者自动化的方式建立,这在很大程度上依赖于现有的词汇知识[6]。如果知识库中没有查询对应的词或者实例,就不可能查到含有它们的文档。因此,获得新词、生成新实例并将它们及时加入知识库中是维护运行信息检索系统的一项重要工作。遗憾的是,目前基于本体的信息检索系统还没有明确提出解决以上问题的有效办法。一方面,由于基于本体的信息检索理论还不成熟,本体论与传统 IR 技术的结合有待进一步研究;另一方面,本体库中的实例包含众多的语义关联,新知识的加入会增加更新程序的复杂度,特别是对于目前以手工维护方式为主的本体存储系统来说,不是一件容易的事情。

3) 语义标记与 HTML 标准不兼容。目前没有统一的标准创建和管理包含 HTML 及语义标注的文档。最常用的方式是将语义标记直接嵌入到 HTML 页面中去,但是考虑用 DAML +OIL 或 OWL 来进行标记时会发现它们是用于知识表示的语言而不是直接嵌入到文本中的。同时在 HTML 页面中嵌入基于 RDF 的标记与 HTML 标准不兼容,W3C 的一个工作组正在研究解决这一问题。

4) 缺乏有效的基于本体信息检索系统的质量评估机制。检索质量评估的目标是对不同搜索引擎系统的检索结果评估其相对优劣次序。目前信息检索领域最重要的评估工作由 TREC 组织负责。TREC 建立了大规模的评估数据集,包括数据集、查询集和相关结果集, 但是 TREC 测试集并不适合基于本体的语义检索系统。测试文档来自专业领域也来自通用领域,并且许多文档带有语义标记,这些都是 TREC 测试集无法提供的。此外,缺乏合理的评估标准对语义标注、基于推理的检索结果以及索引和搜索的性能进行有效的评测。

4 结束语

基于本体的信息检索系统作为本体论与信息检索技术结合的交叉学科领域,已成为国内外学者的研究热点,并取得了许多研究成果。但也应注意到,很多关键技术和问题亟待解决,如针对中文的实体标注技术、实体识别自动工具的开发、本体复用技术、基于软件工程的本体开发方法、本体推理引擎与传统 IR 检索引擎的耦合、自然语言查询优化等。为了开发出实用性强、影响力广的应用项目,基于多媒体信息的本体设计、排序的相关性算法研究、语义服务接口、面向用户兴趣的个性化搜索策略等也是未来研究的热点和发展方向。

参考文献

[1] 顾芳,曹存根.知识工程中的本体研究现状与存在的问题[J].计算机科学,2004, 31(10):1-10.

[2] BAR-YOSSEF Z, KANZA Y, KOGAN Y, et al. Quest: querying semantically tagged documents on the World Wide Web [C]//Proc of the 4th Workshop on Next Generation Information Technologies and Systems. Berlin: Springer,1999: 2-19.

[3] CHINENYANGA T T, KUSHMERICK N. Elixir: an expressive and efficient language for XML information retrieval[J].Journal of the American So ciety of Info rmation Scien ce and Te chnology,2002,53(6): 438-453

[4] 张云中.基于形式概念分析的领域本体构建方法研究[D].吉林大学,2009.

[5] 王晋,孙涌,王璁玮.基于领域本体的文本相似度算法[J].苏州大学学报(工科版),2011,03:13-17+25.

[6] 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,2010,01:51-56.