基于本体的文本信息检索研究

尹蔓卓

（大连海事大学 信息科学技术学院 ， 辽宁 大连）

摘　要:本文对如何构建基于本体的文本信息检索系统进行了探讨，并认为，利用反映概念之间关系的领域本体指导主题标引，利用反映实体之间关系的领域本体指导实体关系标引，并以本体的形式表示文档替代物和查询表达式，可以进一步提高文本信息检索系统的性能。

关键词:本体;信息检索;文本检索;标引

Research of text information retrieval based on Ontology

Abstract: The paper discusses how to construct an on to logy-based tex t inform a tion retrieva l sy stem, and th inkstha t if the sub jec t indexing is based on the dom ain onto logies de scribing the rela tions be tw een concep ts, the entityre lation index ing is based on the dom ain onto log ies de scribing the rela tions between en tities, and the docum en t surrogate s and que ry expressions a re described in the fo rm at of onto logy, the perform ance of the text inform a tion retrieval sy stem w ill be improved.

K eywords: on to logy; info rm a tion re trieval;tex t re trieva l;indexing.

1. 引言：

基于关键词匹配的传统文本检索技术的性能取决于用户对方法的理解，对语义匹配的支持能力较差，具有很大的局限性。尽管基于关键词匹配的检索技术已经有很多次改进，但是由于检索性能没有得到根本的改善，那些没有被文字直接表述出来且隐喻在文中的一些重要信息也就无法被检索。

基于本体的智能检索系统是基于知识的、语义上的匹配，在查全率和查准率上有更好的保证。具体表现为：利用本体，为了消除自然语言理解中的歧义，明确概念含义，在用户提问检索式构造过程中增加语义指导，从而使得构造出的提问检索式能够更加准确地反映用户的真实信息需求；使得用户能够更加准确、方便地实现扩展检索和缩小检索；加强检索系统的推理功能,根据相关概念及背景知识的推理，在完成对信息源搜索的基础上，挖掘出文中的隐含信息,从而实现基于概念的智能检索。综上，基于本体的信息检索将成为一个新的发展方向。

现有的信息检索系统，除了搜索引擎外，大部分系统的信息源都是无结构的文本。这与现有的大部分关于基于本体的信息检索研究不同，现有的大部分检索对象都是Web资源，很少涉及无结构的文本。因此，研究基于本体的文本信息检索依然具有重要的现实意义。

2. 本体概念

本体是一个源于哲学的概念，原意指关于存在及其本质和规律的学说，后来被计算机科学领域引入，特指对共享概念模型所作的明确化、形式化、规范化说明，它强调领域中的本质概念，也强调这些本质概念之间的关联。某个领域的本体能够将该领域中的各种概念及概念之间的关系显性地、形式化地表达出来，从而将概念中包含的语义表达出来。

本体作为一种知识建模工具,自被提出以来就引起了国内外众多科研人员

作者简介：尹蔓卓（1993-）女，辽宁人，硕士研究生 研究方向：数字图像处理

的广泛关注。由于本体能够很好地描述概念以及概念与概念之间的关系,因而将本体引入信息检索系统后,能够为改进信息检索性能提供组织形式和语义上的保证:首先,在信息检索系统中引入本体后,能够最大限度地保留关键词之间的语义关系,大大增强了用户的检索需求表达能力,使信息检索工具更加人性化,查询变得更加方便、直接、有效.此外,基于本体可以进行语义查询扩展,从而检索出与用户查询语义相关的信息;再者,本体通过公理和属性描述概念之间的逻辑关系和规则,提供了对推理的支持。因此,研究引入本体的检索系统意义极大。

在计算机科学领域, 术语“本体”是英文“Onto logy”的中文译法。On to logy 在人工智能或信息系统中的中文翻译, 国内有不同的名称, 如“概念集”、“应用知识体系”,“概念分类体系”,“实体论”,“本体论”、“本体模型”, “本体”、“本体簇”等。由于Onto logy 在英语中的新的含义也是引申来的, 是一个新概念, 所以出现了翻译成不同名称的现象。

四元素表示方法的基本思想是: 一个本体主要有概念、关系、实例和公理这四个元素组成。概念表示某个领域中一类实体或事物的集合, 关系描述概念之间或某个概念的属性之间的关联, 实例是概念表示的具体的事物, 公理用来限制概念和实例的取值范围, 包括许多具体的规则和约束。六元组本体表示方法将本体定义为{C, AC , R, AR , H,X}, 其中C 表示概念的集合;AC表示多个属性集合组成的集合, 其中每个属性集合对应于一个概念;R是一个关系集合;AR是由多个属性集合组成的集合,其中每个属性集合对应于R 中的一个关系;H 表示概念之间的层次结构关系;X 表示公理集合, X 中的元素实际上是概念、关系属性之间的一些约束条件。

3. 本体在信息检索领域的应用现状

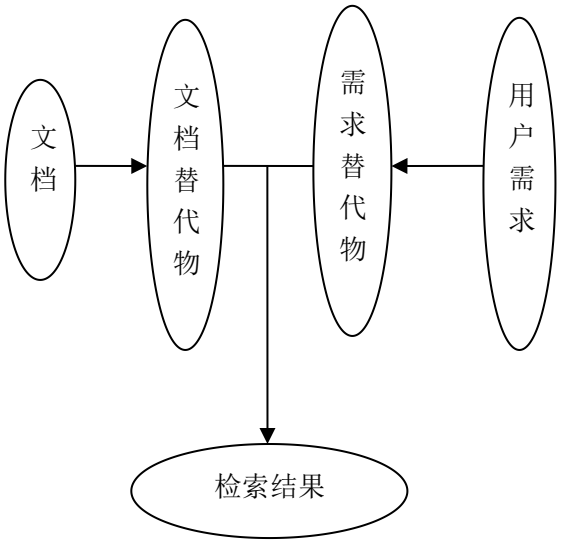
本体是一种技术,在许多涉及知识表示与共享的环境下都可以应用它。 本体之所以能够在信息检索,特别是知识检索中得到广泛应用,是因为本体具有良好的概念层次结构合它支持逻辑推理能力。

基于本体的信息检索的基本思想是:在领域专家的帮助下, 建立相关领域的本体;收集信息源中的数据, 并参照已建立的本体把收集来的数据按规定格式存储在元数据库中;对从用户检索界面获取的查询请求, 查询转换器按照本体把查询请求转换成规定的格式, 在本体的帮助下从元数据库中匹配出符合条件的数据集合;检索的结果经过处理后返回给用户。

基于本体的查询扩展技术最关心的是能否扩展出有效的查询词。尽管目前已经取得了一定的进展,但是由于忽略了属性和实例方面的扩展,扩展出的查询词还是有限的。为了更加准确和全面地反映用户查询意图,需要研究如何基于本体进行和用户查询相关的实例和属性的扩展。

4. 基于本体的文本信息检索系统的一般模型

信息检索过程涉及3 个方面:用户任务(用户需求)、文献的逻辑表示(文档替代物)、相关性判断。为了提高系统的响应速度, 信息检索系统一般不直接进行用户需求和文档的相关性分析, 而是使用表现形式既简单又相似的需求替代物和文档替代物进行相关性匹配, 其一般模型如图1所示。



文档替代物，如关键词向量，可以通过人工标引或者自动标引获得，生成之后一般不再更新;需求替代物，如查询表达式，一般由用户或者检索系统辅助构造而成。文档替代物是文档的元数据，文档类的替代物是文档类的元数据。从检索效率的角度来看，标引能显著提高全文数据库的查全率和查准率，并能缩短检索时间。检索结果可以是题名、知识、概念含义水平上的信息或全文等多种形式。

因为文本内容既包括主题信息，有包含实体关系信息，所以文本标引工作至少应该包含主题标引和实体关系标引两个方面。这是两种不同类型的信息，文本信息检索系统应该提供这两类信息的检索入口。

到目前为止，主题标引技术已经比较成熟，被广泛用于信息检索系统之中，实体关系标引技术还需进一步完善。

信息检索系统引入本体技术的一个重要目的是变关键词(或者主题词) 匹配为基于语义的匹配，使系统在查全率和查准率上有更好的保证。但是，现有的基于本体的信息检索系统知识只是借助于领域本体，判断文档所属领域，对文档按领域进行分类，它在标引的过程中对本体的利用过于简单。领域本体能否在标引过程中发挥更多的作用，有没有必要对标引功能进行相应的改造，例如，使用本体描述文档，用本体作为文档的替代物，这些都有待继续进行研究。

在标引过程中，本体技术可以发挥更多的作用。但是，应该将领域本体分为两类:一类是反映特定领域内概念之间关系的本体，简称概念关系本体;一类是反映特定领域内实体之间关系的本体，简称实体关系本体。这两类领域本体的作用不同，前者用于表达概念体系，只包含单纯的抽象概念之间的关系，例如同义关系、包含关系和实例关系等，相对比较简单。后者就相对比较复杂了，用于一些实在的关系，如，企业之间的兼并关系，合作关系，等等。前者用于主题标引，后者则用于实体关系的标引。

实体关系标引属于信息抽取技术，可视为信息检索技术的一个深化。在标引过程中，实体关系本体可以充当信息抽取框架。实体关系标引的过程可以按以下几个步骤进行:

1.标引系统对文本进行主题标引，识别文本中所包含的主题，并根据文本主题将文本按照领域进行归类。

2.利用文本的领域归属信息从文本中识别出命名实体。

3.利用信息抽取末班将待定的描述信息与实体联系起来。

4.在实体识别是基础之上标注出实体之间的关系。例如职员和组织之间的关系，产品和生产企业之间的关系，以及公司和地区之间的关系，等等。

尽管基于本体的查询扩展方法取得了一定的进展,但许多方法是将查询映射到本体中的概念,或者说,它们所使用的本体更像是简单树形结构的词表,并没有属性和实例概念,能表达的也主要是上下位关系、同义关系,没有考虑到属性、层次、公理和规则的扩展.因而,这样的本体并不能扩展出很多语义关联词.此外,在某些情况下,基于本体的查询扩展方法也可能会减弱某些查询的性能.例如,当新查询词与原查询词的相关性不大时,可能导致“查询漂移”,最终导致“主题漂移”.因此,在将来的查询扩展研究中,一方面要考虑更多的语义关系和从多个角度进行扩展,另一方面应提出有效的查询扩展词推荐策略,对扩展出来的查询扩展词进行过滤,以保证最后得到的扩展词是有效的.

对文本信息检索系统而言，因为文档的多个主题词之间的概念关系比较简单，所以没有必要使用本体形式的文档替代物。文档主题词只是某些概念关系本体上的几个概念节点，用标引词向量作为文档替代物就可以了，但是，需要标注出每个标引词所对应的一个或多个概念关系本体。同样，相应的查询表达式也没有必要采用本体形式。对于W eb信息检索而言，由于信息源以HTM L或者XML语言表达，是半结构化的文本，识别其中包含的各种元数据相对比较容易，元数据之间的关系也比较复杂。

因此，使用本体形式的文档替代物比使

用关键词向量形式的文档替代物更加准确，描述元数据之间的各种语义关系也更加容易。同样，相应的查询表达式也应该采用本体形式。也就是说，基于本体的Web信息检索系统应该以本体的方式表示文档和查询，应该实现基于本体的查询和文档匹配，才能更好地实现基于语义和知识的Web 信息检索。

在精准化和智能化信息检索需求的驱动下,随着本体技术、自然语言处理、机器学习、知识推理等人工智能技术的发展,基于本体的智能信息检索系统已取得一定的进展,但仍然是一个充满问题与挑战的新兴研究领域,可以深入并可能取得成果的方向有很多,主要包括:

1信息检索系统中采用的本体的有效性评估
2新的本体知识的获取和使用
3基于本体的语义标注方法
4基于本体的查询扩展技术
5系统的评测方法
6信息检索系统中的本体匹配和映射机制

5. 结论

随着信息技术与互联网技术的发展,网上的信息资源越来越丰富,造成用户想要找到所需要的信息往往十分困难。信息检索系统被认为可以有效缓解这一难题。

然而，如果采取上述方式构建信息检索系统，那么系统的复杂性、系统实现的难度都将会增加很多，有许多方面的问题需要解决，例如:如何利用领域本体集合对 文档进行标引，生成文档的本体形式的替代物(即文档本体);如何利用领域本体集合生成本体形式的用户查询(即查询本体);如何科学地度量查询本体与文档本体的相关度;如何对本体进行索引，尽可能地保持信息检索系统的响应速度，等等。我们将在以后的工作中进行进一步的研究。

参考文献：

[1]刘肖静，耿骞. Ontology与面向概念的网络信息检索. 情报理论与实践，2004

[2]常春. On to logy在信息管理领域的研究背景. 现代图书情报技术，2003

[3]余一娇. 语义网和语义网格中的本体研究综述.

[4]张红. 语义网中的本体推理及其应用研究.