本体辅助政务信息检索①

屈振新[1]，刘莉萍[2]

[1] (中南财经政法大学 信息与安全工程学院，武汉 430073)
[2] (中南财经政法大学 信息科，武汉 430073)

摘 要: 采用传统基于关键字的全文搜索技术对政务信息进行检索，检索质量比较低。将本体技术与传统检索技术相结合，有利于用户的查询语义充分表达，进而提高检索质量。提出了具体实现方案，以及结合相关领域主题词表来设计政务本体、用本体概念的高频组合词汇提高辅助查询能力的方法。最后通过实例验证了理论的正确性。

关键词: 本体；政务；信息检索

Ontology Auxiliary Government Information Retrieval

QU Zhen-Xin[1], LIU Li-Ping[2]

[1] (School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China)
[2] (Information Department, Zhongnan University of Economics and Law, Wuhan 430073, China)

Abstract: When traditional full-text searching technologies based on keywords are applied on government information, low quality results will be return. It is helpful to express searcher's meaning more completely by combining traditional full-text searching technology with ontology technology. Quality of information retrieval will be improved. A detailed solution is proposed. Later method of designing government ontology utilizing domain oriented thesaurus is discussed. How to use combined words with high frequency to assist searching is also discussed. At last a case verifies that it is right.

Keywords: ontology; government; information retrieval

1 引言

伴随着各种政务活动，产生了大量的公文、文档，当需要某个文档的时候，人们往往迷失在茫茫的信息资源中。采用传统基于关键字的全文搜索技术进行检索，由于处理引擎不能正确理解用户的查询要求，导致检索质量比较低。有些技术通过查找同义词等方法来扩展查询，但扩展的范围有限。

源自于哲学范畴的“本体”是概念模型的明确的规范说明[1]，可以清晰地表达概念间的关系。应用本体技术进行辅助查询，有利于用户的查询语义被充分表达，进而提高检索质量[2]。本文提出了将本体技术与传统检索技术相结合的具体实现方案，利用该方案可以提高政务信息的检索质量。并进一步论述了如何结合电子政务自身的特点设计政务本体，以及用本体概念的高频组合词汇来辅助查询的方法。

2 总体方案

目前的本体语言普遍将描述逻辑作为自身的逻辑基础，描述逻辑中包含了许多构造子，利用这些构造子可以表达概念间复杂的关系。本体不仅能够表达概念间的同义(等价)关系，而且能够表达更丰富、更复杂的关系，如继承和互斥等关系。在本方案的政务本体中，定义了政务领域的基本概念及其关系。

莫里斯和卡纳普认为：句法、语义、语用构成语言的三个基本方面。为了提高检索效率，在电子政务这个特定的环境下，从语用的角度出发，把和每个概

① 基金项目:中央高校基本科研业务费专项资金资助(2009092)
收稿时间:2010-06-07;收到修改稿时间:2010-07-04

念相关的高频组合词汇统计出来，作为本体的补充，辅助查询。

处理过程如图 1 所示：

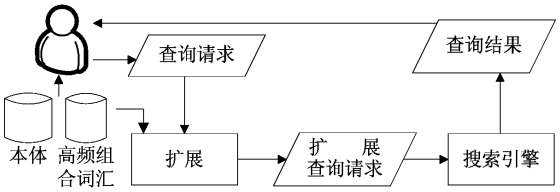

图 1 查询处理过程

用户可以直接提出查询请求，也可以在对本体及高频组合词汇表浏览的基础上，再确定自己的查询内容，提交查询请求。扩展模块结合本体，对查询请求进行语义扩展，包括等价概念的扩展，如果用户要求，还可以自动进行子概念及其它扩展。扩展的查询请求再提交给传统搜索引擎进行检索。

3 政务本体设计

如何设计本体？专家们提出了很多方法，这些方法大多是有工程背景的，是将实际项目开发中的方法、经验总结后得到的通用设计方法。如：斯坦福大学提出了本体设计的七个步骤[3]；METHONLOGY 方法[4]定义了本体设计的三个阶段，它最初是用于指导化学本体设计的，后来也被应用到其它领域本体的设计；TOVE 法[5]是由多伦多大学企业集成实验室提出的，以一阶逻辑谓词为基础，设计过程包括五个步骤；骨架法[6]是用来设计企业本体的方法。

一般来说，设计完全通用的本体不现实，都是针对某个特定的领域来设计领域本体。结合电子政务领域的特点，同时考虑到这里的本体是为信息检索服务的，可以结合电子政务领域现有的信息分类方法，设计电子政务本体。

对于政务信息，国家相关部门发布了多个主题词表，用来对信息分类，便于信息检索。在各类主题词表中，收录了大量规范化的词或词组，称为主题词或称叙词，用于文献的主题标引和检索，而且可以用“用、代、分、属、参”等参照项来描述主题词间的关系，一个主题词和其它主题词之间的关系形成一个网络，在这个网络当中能够表达一定的语义。但是主题词表和本体相比，表现出很多缺点：(1)主题词虽然也是词汇的规范性表达，但是它的逻辑基础不如本体；本体以描述逻辑为逻辑基础，逻辑清晰，表达规范，能够实现对概念和术语的准确表达。(2)主题词间的关系虽然可以用“用、代、分、属、参”等参照项来描述，但是仅能表达简单的语义关系，而且关系的定义并不严格；而本体能够严谨地表达概念间复杂的关系，适合对领域知识进行形式化定义和描述，可以被看作知识库。(3)主题词表在结构、内容上比较保守，一般不经常修改；而本体是开放式的，随时可以根据需求进行调整，适应变化。

虽然主题词表有这些缺点，但是它收录了相关领域的典型词汇，是很有代表性的。所以在设计本体的时候，可以借鉴相关的主题词表，提高效率[7]。可以采取如下步骤：

(1) 明确本体设计的目的和范围。

(2) 收集相关资料。在确定范围内，收集相关资料。包括已经发布的相关主题词表。

(3) 列举出重要词汇。对典型文档、材料进行分析，在主题词表的基础上，扩充相关词汇，得到一个词汇集。

(4) 定义类及其关系。确定词汇集中哪些词汇可以作为类出现，并定义这些类之间的等价、继承和互斥等关系。

(5) 定义类的属性。确定词汇集中哪些词汇可以用来表达类间的关系，将其定义为属性。列举属性可能具有的等价、继承、互逆、传递和对称等特性，并指明基数、取值范围和定义域、值域等限制。

(6) 创建实例。

4 高频组合词汇表的建立

从语用的角度出发，列举出常用的词汇组合，可以增强查询导航的效果，起到更好的提示作用。所以，对于本体中每个概念，找出它对应的高频组合词汇，作为本体的补充。算法如下：

(1) 对选中的样本资料进行词汇抽取处理，得到样本词汇集。

(2) for 本体中的每个概念

在样本词汇集中查找与该概念的相关度不超过 α 的词汇

if 这对词汇共同出现的频率大于 β

将其列为高频组合词汇

相关度：等于文本中两个词汇之间间隔的单词数。

经过该算法计算后，本体中每个概念都可能对应一组词汇，它和其中每个词汇的相关度大，且同时出现的频率高。

如果本体概念的数量为 n，样本词汇集中词汇的数量为 m，算法的时间复杂度为 O(n×m)。

5 实现

为了给全校师生提供高效、智能的公文检索服务，采用上述思想建立了政务信息智能检索系统。

本体的存储、处理使用了 Oracle 11g spatial。

Oracle 11g spatial 支持本体的存储、推理和本体辅助查询，使得开发者能够安全、高效地进行语义应用开发。它在本地实现了 RDF/RDFS/OWL 的支持，可以存储语义数据和本体，支持语义数据查询和关系数据的语义辅助查询，使用自身的或用户定义的推理，以扩充语义数据的查询能力。本体以 RDF 数据模型的三元组的形式存储，Oracle 11g 能够存储数以十亿计的三元组，可以满足大多数应用系统的需求。

所有公文存储在 Oracle 11g 数据库中，公文搜索使用了 Oracle 的全文检索功能。

Oracle 从 7.3 开始就支持全文检索，用户可以使用它的上下文(ConText)选项完成基于文本的查询。它可以在数据库字段上进行全文检索，也支持外部的 doc、pdf 等类型文件，甚至是 URL 文档，可以识别普通文本、XML 及 HTML 等格式的文本。它内含多种词法分析器，支持英文、中文、日文和韩文等多种语言文字，能正确地分词。

(1) 建立本体

按照前述步骤，设计了教育本体。

首先确定了本体是为公文的智能检索服务的，涉及对象是学校公文。经过筛选，选中了 2004 年 304 份典型公文，涉及党务、教学、科研、学生工作、就业和后勤等多个方面。另外，在教育领域，教育部发布了《教育部公文主题词表(试行)》[8]，其中列举了十一个大类，共 228 个主题词，但它只是简单地罗列主题词，没有对主题词间的关系进行定义。

以这两者为基础，和管理人员反复讨论，最后确定了有三百个词汇的词汇集。对词汇集进行归纳整理，确定了学校、教育、人、机构、学校管理、党务、共青团、教学、国际合作、职称、职务、学位、学历、学科、专业、制度法规、科学技术、工程、财务和综合共二十个大类，每个类都有若干子类，如学校下分高校、中等职业学校、中专、技校、中学和小学等，学校管理下分资产管理、后勤管理和医疗管理等。在定义子类时，从不同的侧面来看，可以有不同的定义，要按照最常见的方式来定义。如图 2 所示，用可视化本体编辑器 VOEditor[9]编辑的本体片断：教职工类，可以按照性别，下面包含男职工、女职工两个子类，也可以按照工种，下面包含专业教师、专业技术人员、管理人员和工人四个子类，这两种方法都是可行的，在实际工作中，后面一种提法更多些，所以按照后者进行子类的划分，性别作为类的属性处理。类的层次也不宜过深。比如：人包含教职工和学生两个子类，教职工含专业教师、专业技术人员、管理人员和工人四个子类，专业教师按照职称又可以分为不同的子类，但这样一来，职称就隐含在教职工类中，如果某些地方需要单独使用职称类，就比较困难了，不利于重用，所以人这个类只有三层，职称作为单独的类，利用对象属性将教职工类和职称类联系起来。

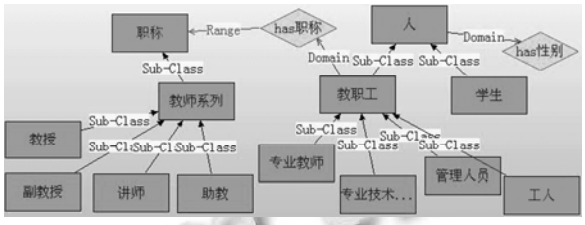

图 2　本体片段

对现有类进行分析，找出相互之间的关系，定义为 ObjectProperty；同时从剩余词汇集中寻找可以用来描述对象特性的词汇，将其定义为相应类的属性。

实例的确定难度比较大，因为同一个词汇在很多情况下，可以被理解为某个类的子类，也可以被理解为某个类的实例，没有一定的标准。这里考虑到本体主要用于导航和智能扩展，只对类进行操作可以简化问题，所以没有定义实例。

(2) 建立高频组合词汇表

对于本体中的每个概念，都列举出与之对应的高频词汇。304 份典型公文存储在 Oracle 11g 中，在相应的表字段上建立 Oracle 全文索引，索引共有四种类型，这里采用 context 类型，索引名为 iexamp。索引建立过程中对正文进行分词处理，得到各个表意单元(也称 term)，一旦索引建立完成，系统自动创建名称为 dr$iexamp$i、dr$iexamp$k、dr$iexamp$r 和 dr$iexamp$n 的表，其中记录了各 term 的出现次数、位置等信息。其中中文 term 有 16495 个，这里只考虑中文 term。term 在所有公文中出现的次数和 term 数的对照表如下：

表 1　term 出现次数和 term 数对照表

出现次数	term 个数
1	9686
2	2436
3	1109
4	673
5	481
6	300
7	227
8	173
9	139
10	137
>10	1134

从表中可以看到，大多数词汇出现的次数很少。出现次数少的词汇肯定不是高频词，为了提高算法效率，选取 term 集中出现次数大于或等于 6 的词汇，作为计算高频组合词汇表的样本词汇。

词汇相关度 α 定义为：两个词汇在同一文档中出现，且间隔词汇数不超过 α 。在 SQL 的 WHERE 从句中，用 contains(……, ……)>0 可以实现这一计算。这里取 α =0。

检索时，Oracle 从表 dr$iexamp$i、dr$iexamp$k、dr$iexamp$r 和 dr$iexamp$n 中查找相应的 term，并计算其出现频率，根据内置算法计算每个文档的得分（score），即“匹配率”。文档中包含相应 term 的记录称为匹配记录，每条匹配记录都有“匹配率”。两个词汇共同出现的频率用所有匹配记录“匹配率”之和表示。 β 的取值如何确定？经过分析发现，有时两个词汇共同出现的频率很高，但没有意义，比如计算本体中的概念“学生”和各样本词汇同时出现的频率，得分最高的是“学生都”，这个组合没有任何意义。所以在确定一个词汇和样本词汇中的哪些词是高频组合时，按照得分降序排列，人工剔除没有意义的组合，取得分排名在前 β 个的组合。这里取 β =10。

结合前述算法，可以计算出本体中每个概念对应的高频词汇表。比如本体中的概念“学生”，经过计算，找出了高频组合：学生工作、学生宿舍、学生管理办法、学生党建、学生干部、学生党员、学生入党、学生人数、学生思想政治工作和学生辅导员。

(3) 检索

用刚建立的本体和高频组合词汇表做辅助，对存储在 Oracle 中的 2530 份公文进行检索。

比如：在这 2530 份公文中，有 5 份公文规定了教师系列职称的评审条件，要查询教师系列职称的评审条件，用“教师系列 and 评审条件”做查询关键字，查询到 9 条记录，其中包含了 5 份正确的公文，查全率=100%，查准率=55.5%；在本体中，“职称”类包含了“教师系列”、“非教师系列”等子类，在“教师系列”下面有包含了“教授、副教授、讲师和助教”子类，用本体对查询条件进行扩展，得到“教师系列 and 教授 and 副教授 and 讲师 and 助教 and 评审条件”，以其作为关键字进行查询，得到 6 条记录，其中包含了那 5 份正确的公文，查全率=100%，查准率=83.3%。

又如：想查询有关学生的公文，输入“学生”之后，因为它是本体中的概念，系统中有其高频组合词汇，自动提示文中常用语“学生工作、学生宿舍、学生管理办法、学生党建、学生干部、学生党员、学生入党、学生人数、学生思想政治工作和学生辅导员”，帮助用户进一步确定查询内容，如果用户认为这不是他关心的主题，可以忽略提示。

6 结束语

本文探讨了如何借鉴相关领域的主题词表建立本体，并在实际应用中进行了验证，但本体设计方法仍然不是工程化的方法，本体的质量和设计者的素质关系密切，如何建立工程化本体设计方法是今后研究的方向。对本体中的每个概念，从样本数据中统计出了它的常见词汇组合，用于辅助查询，如何从样本数据中自动抽取本体还可以做进一步的研究。

参考文献

1 Gruber. A Translation Approach to Portable Ontologies Specifications. Knowledge Acquisition, 1993,5(2):199－220.

2 黄都培.基于本体的法律信息语义检索.计算机工程与应用, 2008,44(28):196－199.

3 Natalya F, Noy and Deborah L. Ontology development 101: A guide to creating your first ontology. [2010-6-1]. http://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

4 Fernandez Lopez M. Overview of Methodologies For Building Ontologies. [2010-6-1]. http://www.lsi.upc.es/~bejar /aia/aia-web/4-fernandez.pdf.

5 Gruninger M. Designing and Evaluating Generic Ontologies. [2010-6-1].http://stl.mie.utoronto.ca/publications/design-generic.pdf.

6 Wand Y, Weber R. Information systems and conceptual modeling a research agenda. Information Systems Research, 2002(4):203－223.

7 薛云,叶东毅,张文德.基于《中国分类主题词表》的领域本体构建研究.情报杂志,2007,3:15－18.

8 教育部. 教育部公文主题词表(试行). [2010-6-1]. http://www2.cqjtu.edu.cn /dzb/show.aspxid=976&cid=12.

9 Li L, Tang SQ, Fang LN. VOEditor: a Visual Environment for Ontology Construction and Collaborative Querying of Semantic Web Resources: Computer Software and Applications Conference. Beijing, 2007. Los Alamitos, Calif: IEEE Computer Society Press, 2007. 591－600.