一种基于本体的语义检索设计与实现

冉 婕，漆丽娟
（云南昭通学院 信息科学与技术学院，云南 昭通 657000）

摘要：基于语义检索的目的，结合成语典故本体的构建，设计了一个基于成语典故本体的语义检索模型，阐述了检索模型中用户界面模块、数据存储模块、查询分析模块和检索分析模块的功能；分析了系统中的本体构建技术、分词技术及检索技术，设计并实现了词语相似度及概念相似度算法；通过语义检索系统的实验，得到较高的查全率和查准率。

关键词：成语典故本体；分词；语义检索；相似度
中图分类号：TN91 **文献标识码：**A **文章编号：**1674-6236(2015)05-0012-03

A design and implementation of semantic retrieval system based on ontology

RAN Jie, QI Li-juan
(*School of Information Science and Technology, Zhaotong University, Zhaotong* 657000, *China*)

Abstract: Based on the purpose of semantic retrieval, this paper designs a semantic retrieval model based idioms ontology, describes the function of the user interface module, data storage module, query analysis module and analysis module in retrieval model. We detailed analysis of its characteristics, design ontology construction technology, word segmentation and retrieval technology in the system. We design and implement the word similarity computing and the concept similarity algorithm. Finally we get a higher recall and precision ratio through the experiment of semantic retrieval system.

Key words: idioms literary quotation ontology; lexical analysis; semantic retrieval; similarity

本体是共享概念模型的明确的形式化规范说明[1]。语义信息检索系统把本体论的技术和方法引入到信息检索系统中，使检索具有一定程度的语义特征，在更高的层次上完成其功能[2-4]。语义信息系统检索的目的就是使机器能够理解信息资源包含的信息内容和用户的信息需求[5]。本系统设计了语义检索系统的总体框架，在所构建的成语典故本体的基础上，融入相似度算法，对基于本体的语义检索相关技术进行了研究。

1 系统的总体框架设计

1.1 ILQO 语义检索系统的设计思路

综合应用 Ontology 的相关方法和信息检索关键技术是实现语义智能检索系统的目标。语义检索的实现实际上就是要将 Ontology 所描述的语义关系应用到对信息资源的检索中，具体就是要通过对本体文件的解析从而在语义层面实现信息检索，并以适当的形式和友好的界面与用户进行交互。成语典故本体的语义检索要解决的关键问题主要有：

1)成语典故本体(ILQO)的创建问题：在 ILQO 的创建中，前期收集了大量的相关资料，并对资料进行整理和分析，借助于 Protégé 这样的辅助工具来实现，目前基于统计学的 Ontology 自动创建技术正在研究之中。

2)Ontology 的存储问题：Ontology 主要是 RDFS 文件格式或者 OWL 文件格式存储。这些文件也可根据标准的格式以 XML 基本语法手工编辑，也可借助 Protégé 等工具自动导出生成，本文是借助 Protégé 工具创建本体并以 owl 文件格式进行存储。此外，本体可以根据需要存储在关系数据库中，本文采取了基于关系数据库的混合存储方式。

3)概念的相似度计算问题：这是实现语义检索的关键步骤，为了更加明确用户的查询请求，本文分别从词语相似度和语义相似度两方面来考虑。词语相似度作为语义相似度的前期工作，主要修正用户提问中的错字、漏字和多字等情况，分别从语素相似度、字序相似度和词长相似度 3 方面来考虑；概念相似度分别从语义相似度和语义相关度两方面来考虑。

4)用户交互界面：为用户提供一个友好的检索交互界面，便于用户查询。这一界面的设计主要是通过 Microsoft Visual Studio C# 技术来设计和实现的。

1.2 系统总体框架

根据上述的设计思路，设计了一个基于 ILQO 的语义检索系统，该系统主要由用户界面模块、数据存储模块、查询分析模块和检索分析模块组成。各个模块相互协作，共同完成检索任务。户界面模块是用户与系统的交互接口，主要处理

收稿日期：2014-06-17 **稿件编号：**201406126
基金项目：云南省教育厅科学研究基金项目资助(2011C040)
作者简介：冉 婕(1975—)，女，四川宣汉人，硕士研究生，副教授。研究方向：智能信息处理。

查询请求及结果回显；数据存储模块实现本体与数据库的转换，用 SQL 来存储本体中的数据；查询分析模块主要对用户的提问进行处理，本系统设计了常见的五种问题模式和答案模式，在检索前对用户问题模式的确定，可一定程度提高语义检索的效率；检索分析模块首先提取问题中的关键词，把关键词作为信息资源的特征项，寻找它在数据库表中的位置，然后根据问题模式对本体库进行检索，最后把检索结果返回给用户界面。

检索分析是语义检索的关键，其主要任务是分析用户的检索目的，即用户想获得的信息，本系统中将检索类型主要分为精确概念的检和语义关系的检索。精确概念的检索是较简单的检索，可直接通过本体数据库获得检索结果，而语义关系的检索则通常是两个或两个以上的关键词且关键词之间存在着密切的语义关系。

综上，给出 ILQO 模型检索的步骤：

步骤 1　对用户输入的查询语句进行分词处理，取出查询中的关键词，进行问题模式的确定，然后将它们递交给检索分析模块。

步骤 2　检索分析模块对递交过来的查询进行分析，并将查询请求分作两种情况来处理：

情况一：精确概念的检索，可直接通过本体库检索结果；

情况二：语义关系的检索，计算两个主概念间的语义相似度，清楚其语义描述，明白用户的检索意图，为检索模块提供了比一般检索方式更准确丰富的信息内容，然后把语义描述交给检索模块，通过对本体库的检索得到检索结果。

步骤 3　综合 1、2 步，将检索结果返回给用户。

2　ILQO 检索模型中关键技术研究

2.1　ILQO

本体构建是语义检索中的一个关键问题。ILQO 构建的具体过程：1）确定本体的应用目的和范围，为了减小本体的规模，本文将本体的范围确定在楚汉相争时期，基于语义检索的目的，建立相应的 ILQO；2）本体分析，定义本体所有术语的意义及其之间的关系，分类是本体构建中非常关键的一步，采用自顶向下的分类法，通过资料的查询，将这一时期的成语典故分作 11 个大类 79 个小类，这种分类方式也便于以后对本体库的扩充；3）领域本体的表示和编码，ILQO 是利用 Protégé 3.2.1 编写完成的，完成后的本体以 OWL 为后缀的文件格式保存。总之，本体建立是对清晰性、一致性、完善性、可扩展性进行检验[6-7]。

2.2　检索中的相似度计算

2.2.1　词语相似度计算及参数的确定

词语相似度计算是概念相似度计算的前期工作，主要是处理用户提问中多字、少字及错字的情况，词语相似度的计算公式如下所示：

$$\mathrm{WordSim}(A,B)=\alpha\cdot\mathrm{CharacterSim}(A,B)+\beta\cdot\mathrm{OrdSim}(A,B)+\gamma\cdot\mathrm{LenSim}(A,B) \quad (1)$$

通过实验测试，文中将公式中的参数 α、β 和 γ 分别取其参考值为 0.7、0.29 和 0.01。通过对本体库中部分概念进行测试，在该参数范围能都能获取较大的相似度值。

2.2.2　概念相似度计算及参数确定

为了实现在 ILQO 上的语义检索，提出了一种基于语义相似度和语义相关度的综合概念相似度计算方法，语义相似度的计算公式如下所示：

$$Sim(C_i,C_j)=\theta_1\times\mathrm{Sim\_dist}(C_i,C_j)+\theta_2\times(\mathrm{Sim\_Info}(C_i,C_j)+\mathrm{Sim\_depth}(C_i,C_j)+\mathrm{Sim\_disenty}(C_i,C_j)+\mathrm{Sim\_symm}(C_i,C_j))/4 \quad (2)$$

其中的 5 个因子分别代表语义距离、信息重合度、深度、密度及不对称因子，θ_1 和 θ_2 为调节参数，且满足 $\theta_1+\theta_2=1$。语义距离的计算如下所示：

$$\mathrm{Sim\_dist}(C_i,C_j)=\frac{\alpha}{\mathrm{Dist}(C_i,C_j)+\alpha} \quad (3)$$

通过实验，公式两个调节参数取值为：$\theta_1=0.8$ 和 $\theta_2=0.2$。

ILQO 系统中概念相似度计算公式如下所示：

$$\mathrm{Sim\_Rel}(C_i,C_j)=\lambda_1\times\mathrm{Sim}(C_i,C_j)+\lambda_1\times\mathrm{Rel}(C_i,C_j) \quad (4)$$

在上面的公式中，实验表明，参数 λ_1 和 λ_2 为 0.5 时获取的相似度值可以较好地体现本体树中不同概念对的重要关系程度（满足：Sim（兄弟）>Sim（父子）>Sim（其他关系）），参数值的改变会出现相似度值相应的改变，但相似度值改变的幅度不大，并且改变参数的不同取值后上述规律仍然满足，故可取 $\lambda_1=0.5$，$\lambda_2=0.5$。

3　系统实现环境及技术

在本节中介绍了系统的实现环境及本体构建和分词技术，通过实验确定了词语相似度计算和概念相似度计算中参数参考值。

3.1　实现技术

基于成语典故的本体构建及语义检索实验系统的开发平台为：

1）Protégé 2.3.1：用于领域本体的创建与维护；

2）RacerPro 1.9.0：领域本体的一致性检测，类层次关系的推理，等价类的推理等；

3）SharpICTCLAS 分词系统 1.0：对用户提问进行分词处理，是关键词提取的前期工作；

4）Microsoft Visual Studio 2008：检索系统的编程、代码设计的主要环境。使用 C#.net 语言完成。

3.2　分词技术

中文词法分析是实现语义检索的基础与关键。首先对用户的提问进行分词处理，抽取其中具有检索意思的关键词，进一步明确用户的检索目的。ILQO 语义检索系统的分词使用的是中科院研制的汉语词法分析系统 ICTCLAS（Institute of Computing Technology，Chinese Lexical Analysis System），它的主要功能包括中文分词、词性标注、命名实体识别等，同时支持用户词典。根据 ILQO 语义检索的特点，ICTCLAS 的词库中缺少相关的词汇，如“成语典故”一词，在 ICTCLAS 的词库中

分别以“成语”和“典故”两个单独的词语出现,不满足本系统的要求,故对其词库 coreDict.dct 进行了扩充,添加了本体中的相关词汇。分词完成后,根据问题模式和答案模式的需求,提取其中的名词作为关键词,具体的分词及关键词的提取如图 1 所示。

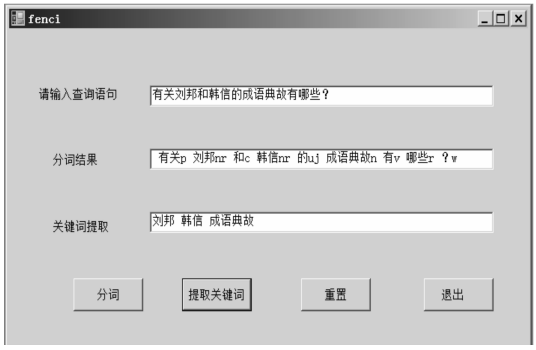

图 1　分词及关键词提取界面

Fig. 1　Extraction interface of lexical analysis and keys

3.3　词语相似度计算

词语相似度计算模型基于 Microsoft Visual Studio 2008 C# 平台实现，程序分别显示 3 个分量相似度参数的取值,计算结果分别显示语素、字序、词长及词语相似度。

4　实例分析

在 ILQO 检索系统中,通过词语相似度的计算,修正了用户提问中的错误,进一步明确用户的检索目的,然后在语义相似度计算的支持下,对语义进行扩展,设计并实现了一个基于 ILQO 和综合语义相似度的文本检索实验系统。基于本体的语义检索系统有领域本体的支持,所以比传统的基于关键字、词的检索更有优越性,检索结果更准确,下面以具体的检索实例进行说明。

例 1:请给出孺子可教的成语典故?

这个检索是简单的检索,我们给出检索的步骤:

步骤 1　对用户的提问进行分词处理,分词的结果为:

请 v 给 p 出 v 孺子可教 n 的 uj 成语典故 n ?w

其中,汉字旁的英文字母是分词中的词性标注。在本检索中,词库虽已有“孺子可教”一词,但其词性标注为 ia,为了提取出我们需要的关键词,故仍对其词库 coreDict.dct 进行扩充,添加“孺子可教”一词,且词性代码标注为 n,相关的算法代码为:

```
string DictPath=@"G:\c# 程序 \SharpICTCLAS 分词系统 1.0\bin\data\";
Console.WriteLine("正在读入字典,请稍候..."+DictPath );
WordDictionary dict = new WordDictionary();
dict.Load(DictPath + "coreDict.dct");
ShowWordsInfo(dict, ´孺´);
Console.WriteLine ("\r\n 向字典库插入 “孺子可教”一词...");
dict.AddItem("孺子可教", Utility.GetPOSValue("n"), 11);
Console.WriteLine("\r\n 修改完成,将字典写入磁盘文件 coreDict.dct,请稍候...");
dict.Save(DictPath + "coreDict.dct");
```

步骤 2　提取关键词,该查询语句的关键词提取为:

孺子可教　成语典故　成语典故

据上,因提取的是分词结果中的名词,该分词结果只有两个名词,故提取后第二、三两个名词相同。

步骤 3　对用户提问进行检索,检索结果如图 2 所示。

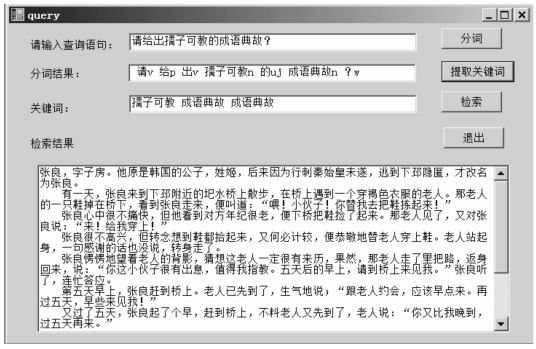

图 2　例 1 检索返回的结果界面

Fig. 2　Results interface of retrieval in case 1

例 2:有关刘邦和韩信的成语典故有哪些?

这个检索是稍复杂的检索,我们给出检索的步骤:

步骤 1　对用户的提问进行分词处理,分词的结果为:

有关 p　刘邦 nr　和 c　韩信 nr　的 uj　成语典故 n　有 v　哪些 r　? w

在本例中,仍然存在对词库的扩充,在此不再赘述。

步骤 2　提取关键词,该查询语句的关键词提取为:

刘邦　韩信　成语典故

步骤 3 在本体库中,计算概念对“刘邦”和“韩信”的相似度,根据相似度的值检索,最终给出检索结果。

分析其检索结果,检索中只有“妇人之仁”一词未检索出,其原因是:在进行知识采集及提炼时,“妇人之仁”一词所涉及到的人物包括刘邦、项羽和韩信三人,而在进行概念相似度计算时,概念对“刘邦”和“项羽”的相似度值为 0.923,而概念对“刘邦”和“韩信”的相似度值为 0.904,前者高于后者,比较下来,该词和概念对“刘邦”和“项羽”的关联性更大,故该词未检索出。

5　结束语

文中论述了成语典故本体的语义检索系统的设计思路,提出了一个基于成语典故本体的语义检索模型,阐述了检索模型中用户界面模块、数据存储模块、查询分析模块和检索分析模块的功能;设计了系统中的本体构建技术、分词技术及检索技术，设计并实现了词语相似度及概念相似度算法。由于自然语言的灵活多变,如何确定特殊问题的模式,如何对本体进行扩充,这将是我们下一步的研究工作。

(下转第 17 页)

据库(表)的建立与维护、数据库(表)的操作,但对数据库而言,还有库(表)的压缩与修复问题,否则,当对数据库(表)进行增、改、删等操作时,将会形成数据垃圾,造成存储空间冗余,ACCESS 也是如此。另外,在 VF 数据库中,包含有各种各样的数据文件,其数据表是独立于数据库之外,这样可以根据文件的信息对数据表进行查询。而在 ACCESS 数据库中,数据表并不独立,是深藏于数据库中的,因为所有的数据表都存储在一个数据库中,有时可能知道数据库的信息,但却不知数据库中数据表及其名称信息。这里，值得注意的是 ACCESS 中的表名在 VF 中变成文件名，操作时需注意将此转换成符合 Windows 命名规则的名称。另外,当文本型数据的字符数大于 254 时,在 VF 中将成为备注型。

3　结束语

除此之外,二者还有很多不同点,如:ACCESS 比 VF 的安全性高,Access 可压缩运行,VF 数据库太大不能压缩运行等等。ACCESS 和 VF 为建立数据库的程序,可以建立一个架构就像是建立仓库,用数据库管理系统可以方便的对房子进行改建、增添;数据库应用系统是对数据库的调用,他若修改数据库要用数据库编程语言经数据库管理系统来修改,就像是你要往仓库里放东西要先去找数据库管理系统才可以。因此,VF 用户在学习 ACCESS 时，可以从以上几个方面理解 ACCESS 和 VF 的差异,在平时的操作中多注意把原先的 VF 知识融汇到 ACCESS 中,就能轻松地学好 ACCESS。

参考文献：

[1] 刘慧. 对VF用户如何学好ACCESS的探讨[J]. 硅谷,2011(20):24-26.
LIU Hui. Discussion onhow userslearnACCESSVF's [J]. Silicon Valley,2011(20):24-26.

[2] 雷景生. 数据库原理及应用[M]. 北京:清华大学出版社,2012.

[3] 何冰. Visual FoxPro环境ACCESS数据库操作的实现探究[J]. 中国-东盟博览,2013(11):10-12.
HE Bing. ACCESS achieve environmental Visual FoxPro database operationsto explore[J]. China-ASEANExpo,2013(11):10-12.

[4] 马铭. ACCESS数据库的安全性 [J]. 吉林师范学院学报,1999(5):33-35.
MA Ming. ACCESS databasesecurity[J]. Journal of Jilin Teachers College,1999(5):33-35.

[5] 王永国. 基于Visual FoxPro环境ACCESS数据库操作的实现[J]. 计算机技术与发展,2011(1):56-58.
WANG Yong-guo. Visual FoxPro environmentto achieve ACCESS database operation[J]. Based onComputer Technology and Development,2011(1):56-58.

[6] 黎升洪. Access数据库应用与VBA编程[M]. 北京:中国铁道出版社,2011.

[7] 李霞林. 任务驱动式教学法在Access数据库教学中的应用[J]. 计算机教育,2006(11):110-112.
LI Xia-lin. Task-driven teaching method in the Access database teaching[J]. ComputerEducation,2011(11):110-112.

[8] 赵启阳,谢锦龙,孙秋分,等. ACCESS数据库使用中"查询过于复杂"问题的解决方案[J]. 石油工业计算机应用,2012(3):44-46.
ZHAO Qi-yang,XIE Jin-long,SUN Qiu-fen,et al. Solution ACCESS database using the"Queryis too complex" problems [J]. Computer Applications of Petroleum Industry,2012(3):44-46.

(上接第 14 页)

参考文献：

[1] Rudi S,Richard Benjamins,Dieter Fensel. Knowledge Engineering: Principles and Methods[J]. Data and Knowledge Engineering,1998(25):161-197.

[2] 马森,赵文,袁崇义,等. 基于规则推理的语义检索若干关键技术研究[J]. 电子学报,2013,41(5):977-981.
MA Sen,ZHAO Wen,YUAN Chong-yi,et al. Research on critical technologies of semantic retrieval based on rule reasoning[J]. Acta Electronica Sinica,2013,41(5):977-981.

[3] 王旭阳,萧波. 基于本体和局部上下文分析的查询扩展方法[J]. 计算机工程,2012,34(7):70-72.
WANG Xu-yang,XIAO Bo. Query expension method based on ontology and local context analysis[J]. Computer Engineering,2012,34(7):70-72.

[4] 王璐,于超,王博,等. 本体语义检索系统[J]. 长春工业大学学报:自然科学版,2013,21(6):56-59.
WANG Lu,YU Chao,WANG Bo,et al. Ontology semantic retrieval system [J]. Journal of Changchun University of Technology:Natural Science Edition,2013,21(6):56-59.

[5] 凌绍东,霍林,王超. 面向语义网的中文本体应用研究[J]. 计算机技术与发展,2014,24(2):194-198.
LING Shao-dong,HUO Lin,WANG Chao. Research on Chinese ontology application oriented semantic network [J]. Computer Technology and Development,2014,24(2):194-198.

[6] 冉婕,孙瑜,昌霞,等. 基于OWL的成语典故本体的构建研究[J]. 计算机技术与发展,2010,20(5):63-66.
RAN Jie,SUN Yu,CHANG Xia,et al. The construction of the idiom story ontology based on OWL[J]. Computer Technology and Development,2010,20(5):63-66.

[7] LI Su-jian. Semantic computation in Chinese question-answering system[J]. Journal of Computer Science and Technology,2002,17(6):933-939.