# 基于语义的成语检索系统研究

傅 鹏[1],涂春梅[1],付春雷[2],马 扬[1],聂奇尉[1]
FU Li[1],TU Chunmei[1],FU Chunlei[2],MA Yang[1],NIE Qiwei[1]
1.重庆大学 软件学院,重庆 400044
2.重庆大学 信息与网络管理中心,重庆 400044
1.School of Software Engineering,Chongqing University,Chongqing 400044,China
2.Information and Network Management Center,Chongqing University,Chongqing 400044,China

**FU Li,TU Chunmei,FU Chunlei,et al.Research on semantic search system for idioms.Computer Engineering and Applications,2011,47(13):147-149.**

**Abstract:** Many current idiom searching systems are constructed with key words matching search pattern.Semantic search system for idioms is proposed to search the desired idiom in the case of users only knowing its meaning without any key character or key word.The implicating concepts in the idiom and the semantic relationships between the concepts are studied, and idiom ontology is created.The searching requirements of the users are analysed from syntax and semantic perspective.Inference engine is used based on description logic for reasoning in idiom area to meet user's searching requirements.
**Key words:** idioms;semantic search;ontology

**摘 要:**现有成语检索系统多采用关键词匹配的检索模式。为了让用户能在仅知道要表达的意思的情况下能够检索到所需成语,提出基于语义的成语检索系统。研究了成语所蕴含的概念和其间的语义关系,构建出成语领域本体,并建造相应的检索系统。该系统首先对用户的查询请求进行语法分析和语义分析,然后对成语领域本体采用基于描述逻辑的推理机进行推理,从而检索出满足用户要求的成语集。
**关键词:**成语;语义检索;本体
DOI:10.3778/j.issn.1002-8331.2011.13.042 **文章编号:**1002-8331(2011)13-0147-03 **文献标识码:**A **中图分类号:**TP391

## 1 引言

成语作为汉语语汇的重要组成部分,存在着丰富的语义关系。现有成语检索系统从本质上讲都是基于关键词的检索。这类检索方式主要是对用户输入的查询请求进行字符匹配,并没有考虑其中的语义关系。如:用户选择通过成语释义来检索成语,分别输入"人品质高尚"、"人的品质高尚"和"人品质好"进行检索,检索出来的结果是完全不同的,然而这三句话实际表达的意思却是相同的。正是由于现有成语检索方式没有对成语以及用户的检索请求进行语义分析,才无法满足用户的这类检索需求。本文针对这一问题提出了基于语义的成语检索系统。

本系统通过构建成语领域本体来表达成语间的语义关系。本体是概念模型的形式化规范说明[1],它能够准确地描述概念与概念之间的内在关系,并通过逻辑推理获取概念之间蕴含的关系,具有很强的表达概念语义和推理的能力。系统通过将用户查询请求中的关键词汇映射到本体的概念中,借助本体的帮助,词汇间就建立了语义关系,再通过推理机进行推理得到相应的检索结果。

## 2 系统框架

基于语义的成语检索系统分为语法分析和语义分析两个子系统,其体系结构如图1。

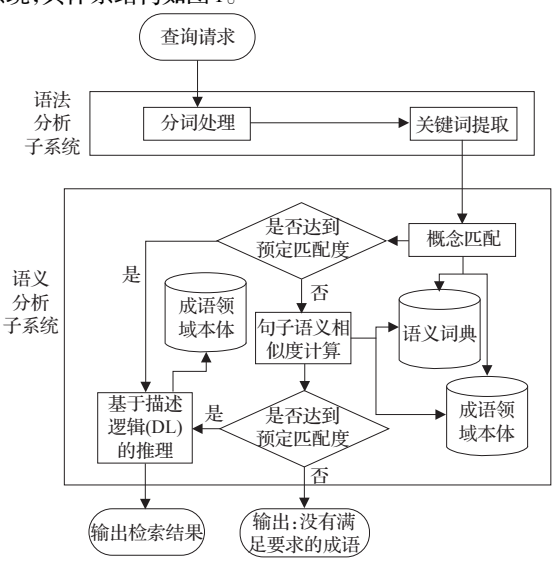


**图1 基于语义的成语检索系统体系结构图**

**作者简介:**傅鹏(1961—),男,教授,主要研究领域为语义技术,信息安全;涂春梅(1985—),女,硕士研究生;付春雷(1978—),男,工程师;马扬,硕士研究生;聂奇尉,硕士研究生。E-mail:fuli@cqu.edu.cn
**收稿日期:**2009-08-10;**修回日期:**2009-11-13

(1)在语法分析子系统中,分词处理模块对用户输入的查询请求进行分词和词性标注。关键词提取模块对整个句子的句法结构进行分析,并结合分词处理后的词性标注提取出关键词,得到相应的关键词序列。

(2)在语义分析子系统中,概念匹配模块是将关键词序列与成语领域本体中的概念进行语义匹配。成语领域本体描述了成语领域中概念及概念之间的各种语义关系,将关键词汇映射到本体的概念中,关键词汇间就建立了语义关系。句子语义相似度计算模块是在关键词序列与本体中的概念匹配未达到预定的匹配度时,通过计算关键词序列与成语领域本体中个体的语义相似度来进行辅助匹配。语义词典是进行概念匹配时进行同义词转换和句子的语义相似度计算的词典资源。基于描述逻辑的推理模块是使用推理机对本体中概念间的语义关系进行推理,从而得到相应的检索结果。

## 3 成语领域本体的构建

成语领域本体是基于语义的成语检索系统的重要组成部分,也是系统进行语义分析的关键,该章主要介绍成语领域本体的构建。本文参照现有的一些构建本体的方法,再结合成语领域的具体情况,使用由美国Stanford大学开发的本体编辑器Protégé4.0来构建成语领域本体。

针对成语领域的一定范围,进行深入的分析总结之后,抽取出其中的相关术语。再对抽取出来的术语进行规范的定义和描述就形成了成语本体中的概念。成语分类在构建成语本体中是一个很重要的问题。本系统对常用成语进行了分析总结后将常用成语分为描写人、事和物三类。由于该本体规模较庞大,图2仅显示了该成语本体的一部分。相关的概念及属性描述如表1和表2所示。

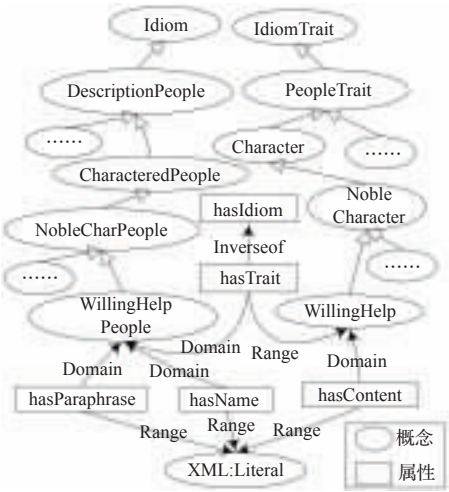


图2 成语领域本体部分图

表1 成语领域本体部分概念表

| 类名 | 中文描述 | 类名 | 中文描述 |
|---|---|---|---|
| Description People | 写人的成语 | Noble Character | 人具有高尚品格这一特性 |
| Description Event | 叙事的成语 | WillingHelp | 人具有助人为乐的高尚品格这一特性 |
| Description Thing | 状物的成语 | Charactered People | 描写人品格的成语 |
| PeopleTrait | 描写人的特性 | NobleChar People | 描写人高尚品格的成语 |
| EventTrait | 描写事的特性 | WillingHelp Peope | 描写人具有助人为乐的高尚品格的成语 |
| ThingTrait | 描写物的特性 | | |
| Character | 人具有品格这一特性 | | |

表2 成语领域本体部分属性表

| 属性名 | Domain | Range | 中文描述 |
|---|---|---|---|
| HasAntonym | Idiom | Idiom | 反义成语 |
| HasSynonym | | | 近义成语 |
| hasIdiom | IdiomTrait | Idiom | 成语所描述对象的特性所关联的成语 |
| hasTrait | Idiom | IdiomTrait | 成语所描述对象的特性 |
| HasParaphrase | Idiom | XML:Literal | 成语所关联的释义 |
| hasName | Idiom | XML:Literal | 成语名称 |
| hasContent | Idiom | XML:Literal | 释义内容 |

图2中成语领域本体的概念描述如表1所示,其中顶层概念包括:Idiom(成语)和IdiomTrait(成语所描述对象的特征)。成语领域本体中概念的层次结构表示了概念之间的继承关系。对于不包含的两个概念之间的关系由表2中的相关属性表示。而每个概念以及概念与概念之间更多更丰富的语义关系还需要对类属性建立约束条件。如所有具有描述WillingHelp特性的成语都是WillingHelpPeope的成语。将这类属性约束条件用描述逻辑(DL)描述如下:

$$WillingHelp \sqsubseteq \forall hasTrait.WillingHelpPeople$$

## 4 关键技术

### 4.1 分词及关键词提取

分词的主要目的是将用户以自然语言提出的查询请求分割成多个词语,并对每个词语进行词性标注。本文使用中国科学院计算技术研究所的汉语词法分析系统ICTCLAS[2]进行分词处理。它的功能包括中文分词;词性标注;未定义词识别等。如,“品格高尚的人”这一句子通过ICTCLAS系统处理后便得到一个带有词性标注的关键词序列:“品格/n高尚/a的/u人/n”。

由于任何句子都是由关键成分(主、谓、宾等)和修饰成分(定、状、补等)构成。关键成分对句子起主要作用。所以只需要考虑句中的关键成分来提取该句子中的关键词。根据语言学知识和对用户输入的查询语句的分析,将句子中的名词、动词、形容词作为关键词,形成关键词序列。将分词后得到的结果(如:“品格/n高尚/a的/a人/n”)进行关键词提取,便得到一个相应的关键词序列(“品格/高尚/人/”)。

### 4.2 语义词典

系统中语义词典的作用主要是对词汇进行同义词转换和计算句子间的语义相似度。本系统使用哈尔滨工业大学信息检索研究室的《同义词词林》作为语义词典。《同义词词林》是现代汉语比较常用的一部类义词典。整个框架是一棵树状层次结构。图3展示了《同义词词林》词义结构的部分示例,每一个节点为一个语义类,所有语义类都表示一个概念,不对应汉语中的词,只有和叶节点对应的语义类是一个具有相似意义的词群。

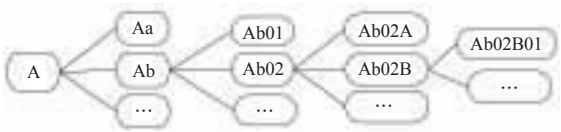


图3 《同义词词林》结构示例图

### 4.3 句子的语义相似度计算

由于成语间丰富的语义关系以及用户查询请求的多样

性,用户的检索请求与概念间的匹配有可能达不到预定匹配度,这就要采用句子的语义相似度计算来进行辅助匹配。通过将用户的检索请求与成语领域本体中描述对象特性的个体(如本体中的概念PeopleTrait的个体)进行相似度计算,将相似度最高且达到了预定匹配度的个体在本体中所属的概念作为概念匹配后的结果。这里的预定匹配度通过实验得到。

具体的句子语义相似度计算方法如下:(1)对句子进行分词处理和关键词提取。(2)计算词语间的语义相似度。本文采用1999年王斌提出的基于《同义词词林》的词语语义相似度计算方法[3]。该方法通过计算词语在同义词词林中的最短路径来计算词语间的语义相似度。(3)计算句子间的语义相似度[4]。设2个句子$A$和$B$包含的词为$A_1,A_2,\cdots,A_m$,$B$包含的词为$B_1,B_2,\cdots,B_n$,则词$A_i(1\le i\le m)$和$B_j(1\le j\le n)$之间的相似度可用$s(A_i,B_j)$来表示。$A,B$句子之间的语义相似度$s(A,B)$为:

$$s(A,B)=\frac{\left(\frac{\sum_{i=1}^{m}a_i}{m}+\frac{\sum_{i=1}^{n}b_i}{n}\right)}{2}$$

式中:

$a_i=\max(s(A_i,B_1),s(A_i,B_2),\cdots,s(A_i,B_n))$

$b_i=\max(s(B_i,A_1),s(B_i,A_2),\cdots,s(B_i,A_m))$

## 5　系统原型

本系统以Java EE技术平台为基础,引入Protégé OWL API软件开发包设计并实现了基于语义的成语检索系统原型。Protégé OWL API是一个Java开源包,它提供类及方法来操作OWL本体。

如图4所示,系统由上至下包括4层。(1)界面层。提供了终端用户的系统访问界面,用户可以通过Web浏览器访问系统。(2)页面服务层。由一些运行在Web服务器上的JSP和JavaBean组件构成。它响应客户端的HTTP请求,根据请求将数据传递给后端的应用逻辑层,并负责将处理结果回送给用户。(3)业务逻辑层。主要由Servlet和Java应用程序组成,它负责处理基于语义的成语检索,根据用户提交的请求对成语本体进行操作。(4)数据层。包括语义词典和成语领域本体。检索时,借助语义词典的帮助在成语领域本体库中进行概念匹配,然后根据概念找到相应的成语个体集合。

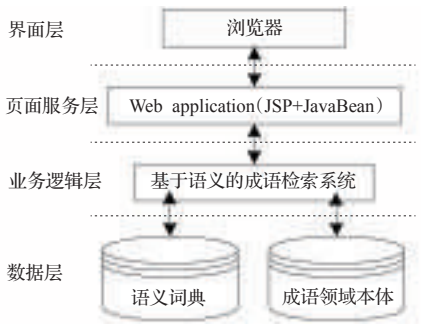


**图4　基于语义的成语检索系统层次结构图**

下面是一个具体的示例。

用户向系统输入检索请求:“一个人愿意帮助别人”。系统首先对用户的请求进行语法分析,得到一个关键词序列:“人/n 愿意/vd 帮助/v”,将其与成语领域本体中的概念进行匹配,匹配结果为:“人/助人为乐”。系统通过调用基于描述逻辑的推理机对概念间的语义关系进行推理就可以得到相应的成语个体集合,检索结果如图5所示。


**图5　检索结果图**

## 6　结束语

针对传统成语检索系统的不足,按照领域信息规范要求,构建出成语领域的本体,使用户可以通过语义检索成语。用户向系统输入查询请求进行语法分析和语义分析后,使用基于描述逻辑的推理机进行推理,得到与用户检索相关的语义信息。下一步工作是要进一步完善成语领域本体,将本体与自然语言处理技术更好地结合,使用户在进行成语检索时能更便捷有效地表达自己的语义查询。同时系统也能快速回应用户的需求,检索出比较满意的结果。

## 参考文献:

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

[2] 刘群,张华平.基于层叠隐马模型的汉语词法分析[J].计算机研究与发展,2004,41(8):1424-1429.

[3] 王斌.汉英双语语料库自动对齐研究[D].北京:中国科学院计算技术研究所,1999.

[4] 秦兵,刘挺.基于常问问题集的中文问答系统研究[J].哈尔滨工业大学学报,2003,35(10):1179-1182.

[5] 陈泳,林世平.基于本体的语义检索技术[J].计算机工程与应用,2006,42(S1):78-80.