

基于成语典故本体的语义检索模型研究

杨雪松 冉 婕

(云南昭通师范高等专科学校计算机科学系 昭通 657000)

摘 要 针对目前信息检索中用户经常找不到或者查不准的情况,提出了一种基于成语典故本体的信息检索模型,并且对该模型及其特性进行了详细分析。在对用户提问进行系统深入分析的基础上,提出了问题模式和答案模式,可提高语义检索的效率,能够在一定程度上提高查准率和查全率。

关键词 成语典故本体,分词,语义检索,相似度

中图法分类号 TP391 文献标识码 A

Research of Semantic Retrieval Model Based on Idiom Literary Quotation Ontology

YANG Xue-song RAN Jie

(Department of Computer Science, Zhaotong Teacher's College, Zhaotong 657000, China)

Abstract Aiming at the problem that the users often can not retrieve useful information or retrieve information is inaccurate, this paper presented an information retrieval model based on Idiom Literary quotation Ontology, and the model and its characteristics were analyzed in detail. Through a systematic in-depth analysis of the users' questions, the questions mode and answers mode were proposed, it can improve the efficiency of semantic retrieval, improve the precision rate and recall rate to a certain extent.

Keywords Idiom literary quotation ontology, Lexical analysis, Semantic retrieval, Similarity

1 引言

以计算机技术、通信技术和网络技术为代表的现代信息技术的飞速发展,改善了信息检索技术的软硬件环境,促进了信息检索理论与实践的发展,同时也给信息检索带来了新的挑战^[1,2]。传统的基于关键词的机械匹配进行检索的方式严重地割裂了字、词间的语义关联,检索过程不包含任何语义信息,导致用户查询获得的检索结果不尽如人意。因此在检索过程中应该导入概念的内容表述——语义,使检索过程由原来的关键词匹配进化为内容匹配,以克服上述仅有表达形式上的匹配所带来的种种缺陷。在这种检索的智能化过程中,本体论扮演了一个十分重要的角色^[3]。由于本体论刻画了事物之间的内在联系,借助于本体论,可以使检索的信息更能满足用户的需求。

本体(Ontology)是“共享概念模型的形式化规范说明”^[4,5]。本体是由特定信息领域中的相关术语集合及这些术语之间的关联所组成的,它的主要目的是对数据库进行描述。本文将以领域本体为基础,研究基于领域本体的语义检索关键技术,并在此基础上设计了一个基于成语典故本体(ILQO, Idiom Literary quotation Ontology)的语义检索模型,为了更加明确用户提问的内容及用户想要知道的答案,提出了问题模式和答案模式,并对该模型的性能和不足之处进行了分析。

本文第2节介绍了成语典故本体语义检索模型的总体设

计;第3节详细介绍了模型中的关键技术;最后总结本文工作并对下一步工作进行展望。

2 成语典故本体语义检索模型

本节详细介绍了 ILQO 语义检索模型的总体设计,阐述了每个模块的功能。

根据上述原理,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、查询分析模块、检索分析模块和数据存储模块组成,如图1所示。

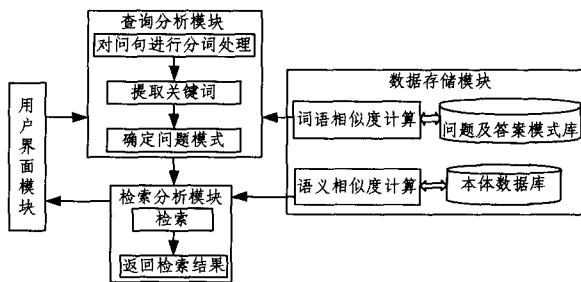


图1 基于 ILQO 的语义检索系统框架

图1中的各个模块相互协作,共同完成检索任务。以下就4个模块分别介绍:

(1) 用户界面模块。用户界面模块为用户提供了与系统交互的接口,接受用户的查询请求,把查询请求提交给查询分析模块去处理,最后返回对查询问题的分析结果,并将其在用户界面显示出来。

(2)数据存储模块。该模块主要将本体中的数据存储到关系数据库中,本着尽量完整保存语义、结构关系清晰的原则,以及检索的方便,设计了概念表、属性表、分类表、实例表、类间关系表和实例关系表 6 个表文件来存储本体中的数据。具体的 ILQO 存储模式如图 2 所示,它给出主要的表及其之间的参照关系。

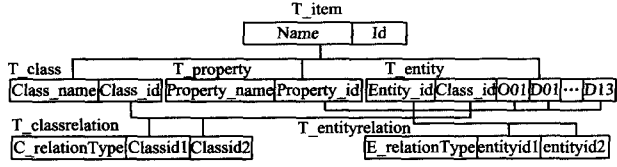


图 2 基于关系数据库的 ILQO 存储模式

(3)查询分析模块。本模块的主要作用是对用户的提问进行处理,首先对用户的提问进行分词,抽取具有检索意义的关键词,然后根据所提取的关键词,结合 ILQO 的特性,对用户提问进行系统深入分析,明确用户提问的内容及用户想要知道的答案,提出了本系统的 5 种问题模式和答案模式,并且问题模式和答案模式是一对一的关系,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。在文章的 3.2 节就问题模式的确定做详细的介绍。系统中的分词部分使用的是 ICTCLAS 系统。根据 ILQO 语义检索的特点,对其词库 coreDict.dct 进行扩充,添加了本体中的相关词汇。

(4)检索分析模块。检索分析模块首先得到查询分析模块提取的关键词,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。

3 ILQO 检索模型中关键技术研究

本节介绍了 ILQO 的特性及建立、ILQO 检索系统中问题模式和答案模式的构建,以及检索过程的分析。

3.1 ILQO

本体构建的设计方法的发展目前处于一个不成熟的阶段,具体的设计方法也不尽相同,比较成熟的方法有:骨架法(Skeletal Methodology)、企业建模法(Enterprise Modelling Methodology)、Methtontology 方法等。这些方法均从各自不同的角度提出了本体建设过程中应该注意的要点^[2]。

在源远流长的中国文化中,成语典故作为古人智慧的结晶、汉语言中的精华,有着言近旨远、形象生动的独有特点。本系统尝试将中国传统文化和计算机技术结合起来,以成语典故语义检索为目的,基于骨架法构建本体的方法,构建了一个成语典故本体^[6](ILQO: Idiom Literary Quotation Ontology),并用本体描述语言 OWL 表示。本体构建过程:①确定本体的应用目的和范围,由于成语典故较多,涉及不同的朝代,为了减小本体的规模,我们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,通过多方面资料的查询,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^[7]。图 3 给出了 ILQO 的片段。

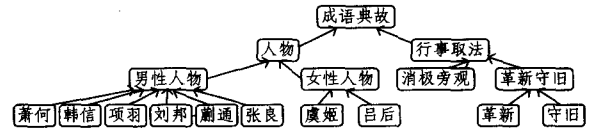


图 3 ILQO 片段

3.2 ILQO 检索系统中问题模式和答案模式的构建

对问题进行分词处理后,可以得到一组词语的组合,过滤其中与领域和问题不相关的词汇,可以得到一组关键词。通过本体中的信息对这组关键词进行分析,可以确定问题的模式。为了确定问题的模式,将本体中的概念分为 3 种类型:主概念、实例和属性。主概念指本体中的分类名,包括子类的名称;实例即本体中的个体,就该本体而言,即各个单独的成语;属性包括本体中的对象属性和数据类型属性。下面分别讨论本模型中考虑到的 5 种问题模式。

(1)模式一:问题中只涉及一个主概念,查询结果是实例。这是该模型中较常见的提问模式。如:和刘邦相关的成语有哪些?

(2)模式二:问题中只涉及一个实例,查询结果是实例的属性值。这是出现频率较高的提问模式。如:多多益善的含义是什么?

(3)模式三:问题中只涉及一个实例,查询结果是实例所属的类别。如:霸王别姬中涉及哪些人物?

(4)模式四:问题中涉及两个主概念,查询结果是两个类下共同的实例。如:有关刘邦和韩信的成语有哪些?

(5)模式五:问题中涉及两个主概念,查询两个类的关系。这种模式是出现概率较小的情况。如:刘邦和吕后的关系如何?

在实践过程中发现名词和动词在句子中起着非常重要的作用,针对本系统的特点发现名词比动词承载着更多的信息量,而且动词的出现基本属于提问的请求词,如“请问”、“请”、“请回答”等。这些动词和查询的主要内容几乎没有直接关系,所以将提问的重心落在名词上,提取关键词时仅提取名词。为了更准确地判断问题的模式,引入了问题模式表和答案模式表,如图 4 所示,两个表联合起来确定问题模式,关键词提取时,分别提取首名词、第二个名词和最末的名词。

问题模式表			
question_id	key1	key2	questionplain
Q01	主概念		查询类下的实例
Q02	实例		查询实例的属性
Q03	实例		查询实例所属的类
Q04	主概念1	主概念2	查询两个类下共同的实例
Q05	主概念1	主概念2	查询两个类的关系

答案模式表			
answer_id	question_id	answer_key	answerplain
A01	Q01	实例	一个或多个实例的集合
A02	Q02	属性	实例的属性值
A03	Q03	主概念	实例所属的类(可能是一个类也可能是多个类)
A04	Q04	实例	两个类下的交集
A05	Q05	属性	类间的关系

图 4 问题模式表和答案模式表

确定问题模式的具体步骤为:

1)对用户的提问进行分词处理,提取其中的名词,形式如:名词1+名词2+...+名词 n ;

2)确定名词1是本体中的哪种类型的概念,再和问题模式表中的key1字段进行匹配;

3)确定名词2是本体中的哪种类型的概念,再和问题模式表中的key2字段进行匹配;若用户提问的名词仅有两个,则名词2为NULL,跳过此步,转到步骤4);

4)确定名词 n 是本体中的哪种类型的概念,再和答案模式表中的answer_key字段进行匹配;

5)综合上面的步骤2)~步骤4):①若匹配成功,则可以确定问题的模式;②若匹配不成功,则进行词语相似度计算后再进行匹配,此时如果匹配成功,那么可确定问题的模式;否则,不属于问题模式库中的模式,不是本系统所考虑的范畴,则提示用户参照本系统的提问模式重新进行提问。

目前在信息检索领域用概念关联度来衡量概念间的联系,从自然语言的角度来讲主要考虑语义相关度和语义相似度两方面因素^[8]。语义相似度是指概念在意义上的相符合程度,在语义树中通过概念的语义距离计算语义相似度,概念的语义距离与语义相似度成反比^[9]。本文采用简单的通过语义距离计算相似度的方法。设有概念 C_1 和 C_2 ,其语义相似度计算方法为:

$$\text{Sim}(C_1, C_2) = 1 - \sqrt{\frac{\alpha - 1}{\alpha} \times \text{Dist}(C_1, C_2)}$$

式中, $\text{Dist}(C_1, C_2)$ 为概念 C_1 和 C_2 间的语义距离, α 是可调节的参数。

3.3 检索分析模块

检索分析的类型主要分为两种:第一种是关于精确概念的检索,主要是对某个特定概念的检索,如问题模式中的前3种;第二种是关于语义关系的检索,这类检索,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,如问题模式中的后两种。

综上所述,我们给出ILQO模型检索的步骤:

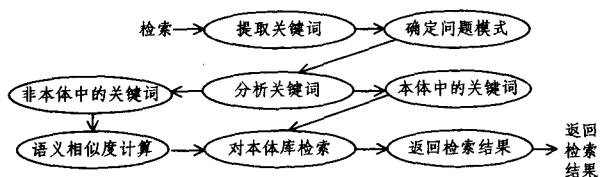


图5 检索分析模块

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 综合上面两步,将检索结果返回给用户。过程如图5所示。

下面以一个简单的实例来进一步说明上述的检索过程。如用户提问:有关刘邦和韩信的成语有哪些?

(1)对用户提问进行分词处理,得到分词的序列:

有关 /pa 刘邦 /nr 和 /ca 韩信 /nr 的 /uj 成语典故 /na 有 /va 哪些 /ra ? /wa

(2)提取其中的名词:刘邦+韩信+成语典故。

(3)名词1“刘邦”属于主概念,名词2“韩信”属于主概念,名词 n “成语典故”属于实例,结合问题模式库和答案模式库,可判定该用户的提问属于问题模式中的模式四。

(4)本实例中的名词都可从本体库中找到与其匹配的概念,故不存在相似度计算,根据前面的分析,对数据库进行检索,最后将检索结果返回给用户。

结束语 针对传统信息检索基于关键词检索出现找不到或查不准的情况,本文提出了一种基于ILQO的信息检索系统,提出了问题模式和答案模式,并将相似度算法融入其中,从而提高了查询的效果。由于自然语言的灵活多变,仍然会影响对问题理解的准确性和完善性,诸如如何确定特殊问题的模式,这将是我们的下一步的研究工作。此外,关于相似度计算的改进也是我们以后需要解决的问题。

参考文献

- [1] 何琳,侯汉清,杜慧平.一种基于领域本体的语义检索系统的设计与实现[J].现代情报工作,2008,52(8):85-88
- [2] 李勇,张志刚.基于本体语义检索技术研究[J].计算机工程与科学,2008,30(4):17-19
- [3] 郑杰,纪希禹.基于本体的语义检索模型[J].科技情报开发与经济,2007,17(25):119-121
- [4] Perez A G, Benjamins V R. Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem Solving Methods [A] // Stockholm V R, Benjamins B, Chandrasekaran A, eds. Proceedings of the IJCAI 99 workshop on Ontologies and Problem Solving Methods(KRR5)[C]. 1999:1-15
- [5] Doan A H, Madhavan J, Domingos P, et al. Learning to Map between Ontologies on the Semantic Web [C] // Proceedings of the 11th International Conference on World Wide Web. New York, USA, 2002:662-673
- [6] 冉婕,孙瑜,昌霞,等.基于OWL的成语典故本体构建研究[J].计算机技术与发展,2010,20(5):63-66
- [7] 韩婕,向阳.本体构建研究综述[J].计算机应用与软件,2007,24(9):21-23
- [8] 万捷,滕至阳.本体论在基于内容信息检索中的应用[J].计算机工程,2003,29(4):122-123
- [9] 刘群,李素建.基于《知网》的词汇语义相似度计算[J].中文计算语言学,2002,7(2):59-76