·数字中医药·

本文引用:刘东波,黄惠勇.基于中医药领域本体的信息检索模型研究[J].湖南中医药大学学报,2017,37(2):220-224.

基于中医药领域本体的信息检索模型研究

刘东波,黄惠勇*

(湖南中医药大学,湖南 长沙 410208)

〔摘要〕针对传统基于关键词匹配的中医药信息检索存在查全率和查准率低下的缺陷,将本体与潜在语义索引相结合,提出一种基于中医药领域本体的语义信息检索模型。该模型基于本体概念扩展树构建相应的查询扩展方法和语义向量空间模型,将用户查询和文档集映射到同一潜在语义空间,通过计算查询向量与文档之间的相似度返回检索结果。着重阐述了该模型的体系结构、实现过程和关键技术,并对其实用性进行论证。

〔关键词〕中医药领域本体;查询扩展;潜在语义索引;信息检索

〔中图分类号〕R2-03 〔文献标识码〕A 〔文章编号〕doi:10.3969/j.issn.1674-070X.2017.02.028

Study of the Information Retrieval Model Based on Traditional Chinese Medicine Domain Ontology

*LIU Dongbo, HUANG Huiyong**

(Hunan University of Chinese Medicine, Changsha, Hunan 410208, China)

〔**Abstract**〕Aiming at the defect of low precision rate and low recall rate of traditional Chinese medicine (TCM) information retrieval based on keywords matching, we propose a semantic information retrieval model based on domain ontology of TCM by combining ontology with latent semantic indexing. Based on ontology concept-extended tree method of query expansion and semantic vector space, the model can map user queries and documents to the same latent semantic space, and returning retrieval results by calculating the similarity between the query vector and the document. In this paper, we focus on the architecture, implementation process and key technologies of the model, and demonstrate its practicability.

〔**Keywords**〕TCM domain ontologyp; query expansion; latent semantic indexing; information retrieval

1 引言

中医学在其长期的发展过程中所形成的医学经典、名家医论、诊疗医案、医学文献具有重要的学术价值和实用价值，是传播中医药知识的重要载体。如何以中医自身的整体观和辨证论治的特点为基础,结合现代信息技术,有效组织、表达和检索中医药信息,已成为总结中医药诊疗规律、转化隐性的中医药诊疗经验为可共享的显性知识、传承和创新中医药知识的必要途径[1]。

当前通过互联网获取中医药信息的途径中,无论是中医药专业网站、医学搜索引擎与目录,还是通用web检索工具(如baidu、Google等),本质上都是基于关键词匹配来获取检索结果。然而,自然语言中所固有的歧义性导致基于关键词的全文检索在对查询的描述上存在模糊性,“一义多词”的存在使得大量相关的信息难以被检索到,“一词多义”的存在使得返回的检索结果中存在大量无关的噪声信息,从而导致检索系统的查全率和查准率低下。

中医学以整体论为指导,采用取象比类的方法,对人体功能状态进行描述,存在大量的古汉语成分,术语描述不规范,一词多义、一义多词的现象普遍存在,数据描述具有模糊性、不确定性和非结构化等特点。这使得中医药知识在客观表达上，在信息的存储、共享和互操作上存在很大障碍,进一步加剧信息检索的不精确性。

〔收稿日期〕2016-10-01

〔基金项目〕湖南中医药大学中医诊断国家重点学科开放基金项目(2013ZYZD17)。

〔作者简介〕刘东波,男,讲师,研究方向:数字中医药。

〔通讯作者〕* 黄惠勇,男,教授,博士研究生导师,研究方向:中医辨证学与数字中医药,E-mail:xuebaozy@126.com。

智能信息检索是支撑下一代互联网的核心技术之一,也是解决中医药信息化过程中高效、准确地获取知识的关键一环。将语义处理技术应用于信息检索,则是智能检索的重要方向。本文将本体与潜在语义索引相结合,提出一种基于中医药领域本体的语义信息检索模型。将本体、自然语言处理、语义向量模型等多种技术相结合,构建基于本体概念扩展树的查询扩展方法和语义向量空间模型,有效提高信息检索的查全率和查准率。

2 系统模型框架

基于中医药领域本体的信息检索模型系统框架如图1所示。该模型将中医药领域本体作为底层知识组织基础,利用本体良好的语义关系和概念层次结构对用户查询请求进行规范化预处理和查询扩展，基于本体构建反映标引词位置权重的词条–文档矩阵和奇异值分解,并对随后的文档集进行语义相似度计算和检索排序。

该模型主要由中医药领域本体、用户界面模块、查询扩展模块、检索分析模块和语义空间模块等组成,各个模块的主要功能描述如下。

2.1　中医药领域本体

在中医专家的指导下,基于中医药领域主题词表及其专业知识来构建领域本体。为了适应领域知识的演化和降低实现的难度,在领域本体的基础上针对所关注的每个细分领域构建相应的分支领域本体。该领域本体将作为语义查询扩展和构建LSI空间的语义基础。

2.2　用户界面模块

接受来自用户的查询请求,并将之提交给查询扩展模块进行处理,最后将排序后的结果集返回给用户。

2.3　查询扩展模块

对用户查询进行分词、去掉停用词等预处理,依据所建立的中医药领域本体将查询词映射到本体中的概念和实例,采用本文所设计的查询扩展方法对查询请求进行语义扩展,并将得到的查询向量提交给检索分析模块进行检索。

2.4　语义空间模块

基于中医药领域本体所生成的概念集商集和个体集商集,对文档资源库中的原始文档构建语义索引和k维LSI语义空间。

2.5　检索分析模块

将查询向量映射到LSI语义空间,将之与文档集中的每个文档进行语义相似度计算,选取语义相似度大于给定阈值的文档,并在排序后作为结果集返回给用户。

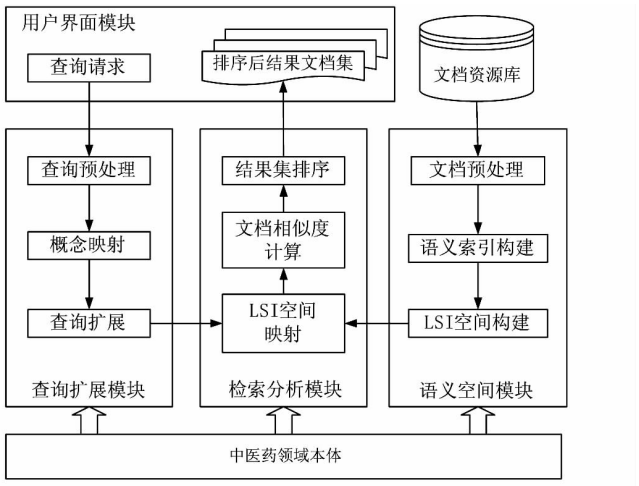

图1 基于中医药领域本体的语义信息检索模型框架

3 关键技术研究

3.1　中医药领域本体构建

本体是共享概念模型的形式化规范说明[2],它能确定相关领域内共同认可的术语，提供对该领域知识的共同理解。中医药本体能够对中医药领域的概念以及概念之间的联系进行规范化描述,消除中医药知识的模糊性和不确定性，在语义层面建立该领域的共享概念模型。利用中医药领域本体的形式化描述能力，可以消除中医药术语中普遍存在的一词多义、一义多词所带来的知识表达的不确定性,促进领域知识的共享和重用,使得患者、医护人员和科研人员能够快速准确地获取所需的中医药信息。

为了降低实现的难度，本文在分析中医药知识结构和常用概念的基础上，针对中医药领域中的中医诊断这一分支领域构建中医诊断本体。以《中国中医药主题词表》[3]、GB/T 16751.2–1997中医临床诊疗术语(证候部分)、《中医诊断学》[4]等内容作为数据平台,以专用于领域本体构建的七步法[5]作为本体构建方法,在中医专家的指导下,借助于part–of、kind–of、instance–of、attribute–of等4种基本概念关系[6](见表1),对中医诊断当中的概念及其关系进行描述,构建中医诊断本体，并由此得到一棵反映该领域知识层次结构特点的概念层次树。由于这种层次结构并不要求子女仅有一个父节点，以及attribute–of关系的存在,因此它实际上是一个有向无环图。见图2。

由于名词（或者一组名词）最能代表文档的内容，因而使用中医诊断本体中所定义的概念和个体来标记文档,并据此构建文档逻辑视图。下面给出建立文档逻辑视图所需的两个定理，其证明过程参见文献[7]。

定理1:设本体定义的所有概念的集合为S(con-

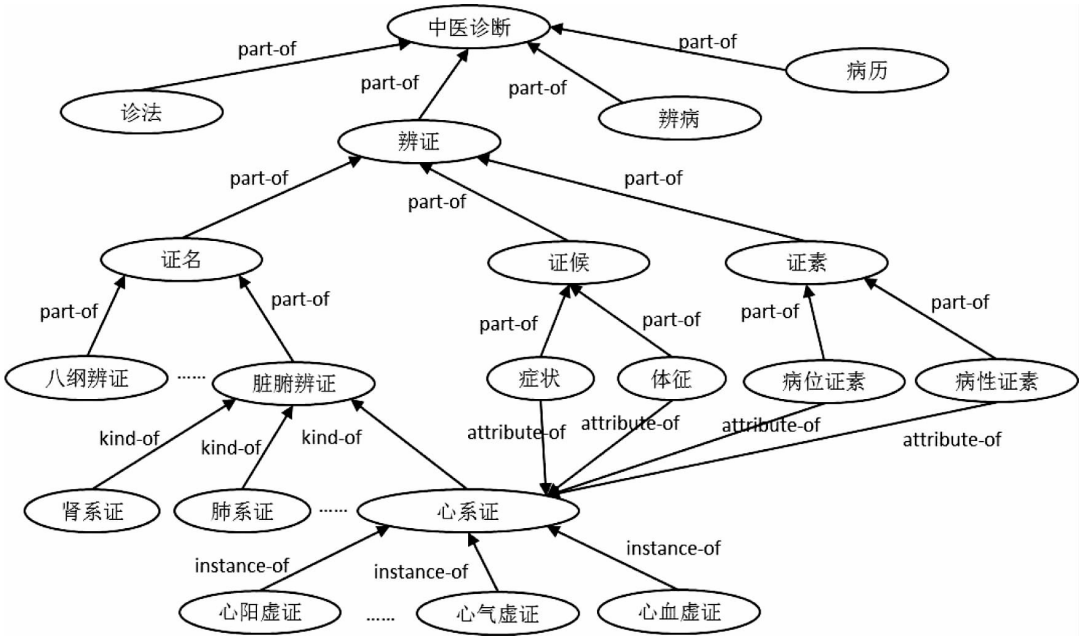

图2 中医诊断本体片段

表1 本体4种基本概念关系含义

关系名称	含义	举例
part-of	表达概念之间部分与整体的关系	“证候”与“症状”
kind-of	表达概念之间的继承关系	“脏腑辨证”与“心系证”
instance-of	表达概念的实例与概念之间的关系	类“心系证”与实例“心气虚证”
attribute-of	表达某个概念是另外一个概念的属性	“症状”是“心系证”的一个属性

cept),根据两个概念间的等价关系(≡),利用*tableau*算法可建立商集S(concept)/≡。

定理2:设本体定义的所有个体的集合为*S(individual)*,利用本体中已有的关于个体间的等价关系,反复调用个体间等价关系*EI*满足的3条规则,根据个体间等价关系*EI*,可建立商集*S(individual)/EI*。

合并中医诊断本体中的概念集合和个体集合,得到本体词汇集合*M*,即*M=S(concept)*∪*S(individual)*,由定理*1*和定理*2*可分别推导得出相应的商集,将之合并形成一个语义索引项集合{[X_1],[X_2],…,[X_t]}(设为*I*),则有*I*=S(concept)/≡∪*S(individual)/EI*。令标引词集合*I′*={X_1,X_2,…,X_t},并根据*I*和*I′*对用户查询进行语义扩展和建立基于本体的标引词-文档矩阵。

3.2 基于本体的查询扩展机制

从扩展词来源角度,可将查询扩展技术分为全局分析、局部分析和基于关联规则的查询扩展技术等几种。这些技术由于没有关注查询词之间的语义关联,因而存在查询歧义性问题,不能消除用户查询与检索结果之间的语义偏差,并且由于大量无关词加入扩展集合,容易产生“查询漂移”问题[8]。

为了充分表达和扩展用户查询意图,本文将用户查询词与中医诊断本体中的概念和个体进行匹配和概念扩展。由于同义、继承以及整体-部分关系是概念之间关系的主要组成部分,因而本文主要针对这些关系进行语义查询扩展。

在对查询请求进行划分语句、分词与词性标注、去掉停用词等预处理操作后,采用文献[9]提出的概念词典,将查询请求转化为初始查询向量(设为q)。概念词典包含有语法和词汇信息,具有词汇与概念之间的相互映射,通过它能使词汇迅速被抽象为概念。

依据语义索引项集合*I*,合并向量q中的等价元素,并用标引词集合*I′*中对应的标引词予以表示,由此得到查询向量$q'=(q_1,q_2,\cdots,q_s)$。对于查询向量q'中的两类分量,采用不同的语义扩展策略。当查询词(查询向量q'的分量)为概念时,依据继承关系和整体-部分关系向下扩展*l*层得到一棵概念扩展树,将扩展树中的概念及其个体加入查询向量,此时依据继承关系扩展得到的扩展概念节点的权重为δ^l (δ为常数,$0<\delta<1$),依据整体-部分关系扩展得到的扩展概念节点的权重为θ^l(θ为常数,$0<\theta<\delta<1$),从概念节点扩展到其个体节点的权重为ε (ε为常数,$0<\varepsilon<1$);当查询词为个体时,为了避免“查询漂移”问题,只将该查询词的直接概念向下扩展一层即可,此时扩展

所得查询词的权重仍然采用前面的策略。经过以上步骤,得到一个最终的查询向量$q^*=(q^*_1,q^*_2,\cdots,q^*_s)$及其权重向量$w=(w_1,w_2,\cdots,w_s)$。

3.3　基于本体的LSI语义空间

为了避免仅采用标引词(*index terms*)表示用户查询和文档所带来的性能缺陷(较低的查全率和查准率),本文基于中医诊断本体构建标引词–文档矩阵,以此建立基于本体的潜在语义索引空间。

由于词典标引法 [10] 在汉语自动标引中广为使用,本文采用该法以实现文档的自动标引,并据此建立本体词汇–文档矩阵$[b_{ij}]_{m\times n}$(设为B,通常为稀疏矩阵),其中:

m:文档库中不同的本体词汇数;

n:文档库中的文档数;

b_{ij}:第i个本体词汇在第j个文档中出现的频度。

由于文档中存在从属于本体词汇同一等价类的词条,因而需要将它们当做一个整体进行词频权重计算,方法如下:

令$u^T_1,u^T_2,\cdots,u^T_m$为矩阵B的行向量组,依据索引项集合$I(I=\{[X_1],[X_2],\cdots,[X_t]\})$,合并向量组$u^T_1,u^T_2,\cdots u^T_m$中属于同一等价类的向量组$u^T_{p_1},u^T_{p_2},\cdots,u^T_{p_r},(p_1,p_2,\cdots,p_r\in[1,m])$,由此得到一个新向量,其对应的标引词为该等价类的标引词,其每个元素值为向量$u^T_{p_1},u^T_{p_2},\cdots,u^T_{p_r}$中各对应位置元素值的代数和:

$$A=[a_{ij}]_{t\times n} \tag{3-1}$$

其中:

t:文档库中包含在集合$I'(I'=\{X_1,X_2,\cdots,X_t\})$中的标引词个数;

n:文档库中的文档数;

a_{ij}:所有与标引词X_i等价的本体词汇在第j个文档中出现的频度代数和。

为了提高语义空间的质量,还需对词频矩阵进行加权处理,通常采用局部加权策略和全局加权策略来分别评价标引词在某个文档和整个文档集中的重要程度,其形式如下:

$$a_{ij}=L(i,j)*G(i) \tag{3-2}$$

其中,$L(i,j)$表示标引词i在文档j中的局部加权函数,$G(i)$表示标引词i在整个文档集中的全局加权函数。目前在LSI中广泛采用的TF–IDF加权策略基于香农信息学理论:1)如果词条在所有文档中出现的频率越高,则其所包含的信息熵越少;2)如果词条只在少量文档中拥有较高的出现频率,则其拥有较高的信息熵。以下是对TF–IDF进行加权的最著名的方法[11]:

$$L(i,j)=tf_{ij},G(i)=idf_i=\log(N/df_i) \tag{3-3}$$

其中,tf_{ij}表示标引词t_i在文档d_j中出现的频度,idf_i表示标引词t_i反比于标引词出现的文档频度,N表示文档集中全部文档的数目,idf_i表示含有标引词t_i的文档数目,$\log(N/df_i)$也称倒排文本频率。

对已经建立的标引词–文档矩阵A进行奇异值分解,得到包含A的左右奇异向量的正交矩阵U和V,以及奇异值对角矩阵S:

$$A=USV^T \tag{3-4}$$

其中,$S=diag\ (s_1,s_2,\cdots,s_r),rank\ (A)=r\leqslant p=\min(t,n),s_1\geqslant s_2\geqslant\cdots\geqslant s_r>0$。

选取S中前k个最大的奇异值,并选取U和V中相应的行和列,由此得到A的k–秩近似矩阵$A_k[k<<\min(t,n)]$:

$$A_k=U_kS_kV^T_k \tag{3-5}$$

此时,U_kS_k的行可看作标引词在k维LSI语义空间中的向量,$S_kV^T_k$的列看作是文档在语义空间中的向量。由于S_k为对角矩阵,对k为空间中的坐标进行适当缩放即可用U_k代替U_kS_k,因而可以将U_k中行向量看作标引词向量。同理,可将V^T_k中列向量看作文档向量。

通过奇异值分解和k–秩近似矩阵,单个的标引词被导出的正交因子替代,极大的降低标引词之间的“斜交”现象,消除标引词–文档矩阵中包含的“噪声”信息,从而更加凸显标引词与文档之间的语义关系。另外,基于本体等价类词汇构建标引词–文档矩阵和截取k–秩近似矩阵,极大地减少标引词与文档向量空间,提高文档的检索效率。

3.4　查询匹配与语义相似度

为了将用户查询和已被映射到语义空间中的文档进行相似度计算,需要首先将用户查询向量映射到k维LSI语义空间。Tamara在文献[12]中的实验表明,用户查询映射过程中是否使用局部权值,对于检索结果没有明显影响。另一方面,为了体现标引词的全局区分度,则有必要对用户查询的标引词应用同样的全局权值。为此,根据标引词–文档矩阵A中标引词的排列顺序和在查询语义扩展阶段得到的查询行向量$q^*=(q^*_1,q^*_2,\cdots,q^*_s)$及其权重行向量$w=(w_1,w_2,\cdots,w_s)$,本文提出如下公式以得到用于语义映射的列向量$q[(q^T=(q_1,q_2,\cdots,q_m)]$。

$$q_i=\begin{cases}0, & w_i\notin w\\ w_i*\log(N/df_i), & w_i\in w\end{cases} \tag{3-6}$$

然后,根据公式(3–7)得到查询向量q在LSI中的向量表示:

$$\hat{q}=q^T U_k S^{-1}_k \quad (3-7)$$

其中，$q^T U_k$等价于将q映射到LSI向量空间中，右乘S_k^{-1}则是赋予LSI向量空间每一维不同的权重。最后，采用向量夹角的余弦值来计算$\hat{q}$和文档向量之间的相似度：

$$sim(\hat{q},s_j)=\frac{\hat{q}\cdot s_j}{|\hat{q}\|s_j|} \quad (3-8)$$

其中，s_j表示文档d_j在$S_k V^T_k$中对应的第j个列向量，也即文档d_j在LSI语义空间中的向量表示，$\hat{q}\cdot s_j=\sum_{i=1}^{k}\hat{q}s_{ji}$，$|\hat{q}||s_j|=\sqrt{\sum_{i=1}^{k}\hat{q}^2_i}\cdot\sqrt{\sum_{i=1}^{k}s^2_{ji}}$。降序排列相似度的计算结果，并将相似度大于阈值$S_{\min}$的文档作为检索结果$D_{Result}$返回给用户：

$$D_{Result}=\{d_j|\mathrm{sim}(\hat{q},s_j)\geqslant S_{\min}\} \quad (3-9)$$

4 模型实用性分析

由于目前还没有用于基于本体的智能信息检索系统的标准评测框架，已有的一些评测方法都是以用户为中心的，不可扩展，很难复用[13]，因此本文从定性分析的角度来验证结合本体与潜在语义索引的信息检索模型的实用性。

宋峻峰等[7]通过形式化定义和推理，从理论上验证基于本体的信息检索模型具有比传统信息检索模型更好的文档逻辑视图和用户信息需求逻辑视图。本文通过part-of、kind-of、instance-of、attribute-of等4种基本概念关系对中医诊断当中的概念及其关系进行描述，构建中医诊断本体，为该领域的知识共享和共同理解奠定基础。本文所提出的语义信息检索模型在文献[7]的基础上，进一步考察中医诊断本体概念之间关系中的占绝大部分比例的同义、继承以及整体-部分关系，构建基于本体概念扩展树的查询扩展方法，使得查询条件更加符合用户意图，有效的降低用户查询意图与检索结果之间的语义偏差。

传统向量空间模型采用标引词来表示用户查询和文档，基于TF-IDF加权策略来计算标引词在文档中的权值和构建标引词-文档矩阵。由于语言中一词多义、一义多词的普遍存在，导致出现影响向量空间模型检索性能的标引词“斜交”现象。潜在语义索引模型通过对标引词-文档矩阵进行奇异值分解和取k-秩近似矩阵，极大的降低了标引词之间的“斜交”现象，更加全面的再现标引词和文档之间的关系，克服单纯项表示时产生的同义、多义以及“斜交”现象，提高了检索的查全率和查准率[14]。另外，由于采用本体等价类词汇作为标引词，以及采用k-秩近似矩阵（k远小于标引词-文档矩阵维数），极大地减少标引词与文档向量空间，有效提高检索效率。

5 结语

针对传统基于关键词匹配的中医药信息检索在查全率和查准率方面的局限性，特别是在信息语义表达上的不足，本文提出一种基于中医诊断本体和潜在语义索引的中医药信息检索模型。通过构建中医诊断本体，提供对该领域概念以及概念之间联系的规范化描述，并据此构建查询扩展方法和潜在语义空间，使得该模型比传统信息检索模型更好的反映文档和用户信息需求语义，有效消除中医药知识的模糊性和不确定性，有助于提高患者、医护人员和科研人员获取所需中医药信息的准确性和完备性。在下一阶段的研究中，我们将在中医专家的指导下进一步完善中医诊断本体，并通过大规模数据集测试和优化本信息检索模型。

参考文献：

[1] 谢 琪.基于本体方法构建中医药概念信息模型的方法学示范研究[D].北京:中国中医科学院,2011.

[2] Borst WN. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. Enschede, University of Twente, 1997.

[3] 吴兰成.中国中医药主题词表[M].北京:中医古籍出版社,1996:85-98.

[4] 朱文锋.中医诊断学[M].2版.北京:中国中医药出版社,2007:139-215.

[5] 陆建江,张亚非,苗 壮,等.语义网原理与技术[M].北京:科学出版社,2007:70-73.

[6] 邓志鸿,唐世渭,张 铭,等.Ontology研究综述[J].北京大学学报(自然科学版),2002,38(5):730-738.

[7] 宋峻峰,张维明,肖卫东,等.基于本体的信息检索模型研究[J].南京大学学报(自然科学版),2005(2):189-197.

[8] 欧阳柳波,谭睿哲.一种基于本体和用户日志的查询扩展方法[J].计算机工程与应用,2015,51(1):151-155.

[9] 李振东,费翔林.基于概念的信息检索模型研究[J]. 南京大学学报(自然科学版),2002(1):99-109.

[10] 王知津.现代索引文摘法[M].北京:北京图书馆出版社, 1999.

[11] 李媛媛,马永强.基于潜在语义索引的文本特征词权重计算方法[J].计算机应用,2008,28(6):1460-1462.

[12] Tamara GK. Limited-Memory Matrix Method with Applications, Doctor's dissertation, University of Maryland, College Park, 2001.

[13] 杨月华,杜军平,平 源.基于本体的智能信息检索系统[J].软件学报,2015,26(7):1675-1687.

[14] 林鸿飞,姚天顺.基于潜在语义索引的文本浏览机制[J].中文信息学报,2000,14(5):49-56.

（本文编辑 李 杰）