中医药文献语义关系图发现

陶金火[1] 陈华钧[1] 胡雪琴[2]
(浙江大学计算机学院 杭州 310027)[1] (中国中医药科学院 北京 100700)[2]

摘 要 提出了一种基于中医药语义本体知识库对中医药文献进行语义关系图发现的方法。核心方法分为三个部分:第一步采用中医药语义本体概念名称为字典进行关键词提取;第二步采用关联算法的一种变异算法查找高频关键词组;第三步利用中医药语义本体知识库对关键词组进行语义关系识别,对未能识别的关键词进行语义关系预测。最后每组关键词生成一个对应的语义关系图。实验部分将利用中医药语义本体知识库对中医药文献进行语义关系图的发现,验证提出的算法。

关键词 中医药语义本体,语义关系图
中图法分类号 TP39 **文献标识码** A

Semantic Graph Discovery of TCM Documents

TAO Jin-huo[1] CHEN Hua-jun[1] HU Xue-qin[2]
(Department of Computer Science, Zhejiang University, Hangzhou 310027, China)[1]
(Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China)[2]

Abstract This paper proposed an ontology based TCM semantic graph discovery of TCM Document. The core method includes three procedures. Firstly, extract keywords from the TCM documents, using the TCM ontology concept name as dictionary. Secondly calculate the frequency of the keywords. Thirdly, identify the semantic relation between the keywords with the TCM ontology knowledge base. Furthermore, predict the semantic relation that can't be identified. Therefore, every group of keywords could generate a semantic graph that express the possible semantic of the original sentence. In the experiment section, the TCM ontology knowledge base was used to identify the semantic graph from TCM Documents, and verify the feasibility of the method of this paper.

Keywords TCM ontology, Semantic graph

1 引言

历史悠久的中医药领域拥有大量的包含宝贵知识的文献,对中医药文献的自动化的分析处理一直是学界的研究热点。不过受限于中医药文献的一些特点,比如中医药学的概念表达、术语使用甚至语句内容都具有特殊领域性,计算机对这些文献的分析处理一直不太理想。语义 Web 技术作为下一代互联网技术,将人类的所有知识进行无缝链接集成,为中医药文献的分析处理提供了新的解决方案。

语义 Web 的概念最早由 T. Berners-Lee 与 J. Hendler 等提出[1,2]。语义 Web 的主要目标是在 Web 中增加机器可以理解的语义,提升机器对 Web 内容的理解,从而更好地支持搜索内容、知识发现、内容推荐等服务。语义 Web 通过语义关系图来表示现实世界的所有事物之间的语义关系,从而构成一个巨大的知识和资源网络,对应用领域提供各种相关的服务。语义 Web 的核心概念是语义本体(semantic ontology),本体可以涵盖任何概念模型。

作为语义 Web 的基础,语义图(semantic graph)以语义本体为节点,以 RDF 三元组〈主语,谓词,宾语〉的形式描述节点之间的语义关系,也即图的边。语义图可以表达任意复杂的语义结构、各个元素及各元素之间的语义关系。语义图是一种先进的知识表达模型,具有直观性、全面性、可推理性等特点。语义图通常用 RDF 或者 OWL 来表述,使计算机能够很好地识别,便于自动化分析处理。将语义图应用于中医药领域,进行知识的表达和建模,可以使中医药知识变得统一规范,有利于人们对中医药知识的获取、存储和分享,同时增强自动化分析处理的能力。

语义图在处理中医药文献方面能发挥很大的作用。语义关系图能够以短语、句子、段落等为单位对其内容进行语义建模。比如:对于一个有主语谓语宾语的句子,主语和宾语用语义图的节点来表示,而谓语则用顶点之间的有向边来表示。文献可以以句子为单位进行语义图构建,所有的子图按照一定规则进行连接,从而实现对整个文献的建模。另外如果将各个节点归并到其所属的直接父节点或者更上层的节点,就可以得到精简的概括性的图,这个图概括性地描述了文献的内容,可以用于对文献的分类;对语义图进行节点频度分析、知识推理等,可以得到文献的主题词。总之,将文献转化为语义关系图,对文献的分析处理文献知识提取、文献主题词标

到稿日期:2010-04-21 返修日期:2010-07-25 本文受 NSFC61070156,2009QNA5025,2010QNA5044 资助。
陶金火(1985—),男,硕士生,主要研究方向为语义知识发现等,E-mail:taojinhuo@gmail.com;**陈华钧** 男,副教授,主要研究方向为网格计算、语义 Web 等;**胡雪琴**(1978—),女,助理研究员,主要研究方向为中医药信息学等。

引、文献分类等方面都有着重要的意义。本文将重心放在从中医药文献中对语义关系图的提取，提出了一种利用中医药语义本体知识库，从大量中医药文献中，进行语义关系图发现的方法。在实验部分本文用中医药本体知识库对中医药文献进行语义关系图发现，以验证算法。

2 相关工作

Leskovec J 等提出一种对文档建立语义关系图的方法[8]。他们的方法是对文档进行深度的语法分析，对每个句子生成一个主语谓语宾语的三元组，并在此基础上，生成文档的语义关系图。不同于利用语法分析，本文提出的方法采用领域字典提取关键词，因为中医药文献文法的复杂性，目前还没有很好的自动化的语法分析方法，另外采用领域字典分词也提高了关键词提取的精度。

Xiaogang Zhang 提出了一种对中医药文献进行二元语义关系的挖掘方法[9]。其利用 TCM 本体信息，对中医药文献进行二元语义关系发现。本文方法是对文献语义关系图的挖掘，将获得一种更全面的语义关系，包含更全面的文献的知识，语义关系图的作用在引言部分已经表述过了。

Hasegawa Takaaki 等提出了一种将实体对进行聚类分析来抽取语义关系的方法[11]。该方法的优点是无须事先标注训练数据，但是其分类结果的准确性不高。本文提出的语义关系图发现的方法利用已存在的本体数据作为知识库，对文献进行语义关系图的发现，在准确率和覆盖率上都较好。

关联规则算法(Association Rules Mining)是数据挖掘中一种重要的分析方法[4]。关联规则挖掘是为了在数据库中发现关联关系，是数据挖掘最先研究的问题之一。本文结合中医药文献分析的具体情况，采用关联规则算法的一种变异算法进行高频关键词的计算。主要改变是项集中的元素具有一定顺序和范围的要求。

向量空间模型最早由 G. Salton 和 A. Wong 提出，用于信息的索引和获取[7]。现在其广泛地应用于文档的分析处理，包括文档信息、文档特征项、特征项权重等的表示。文本对文献进行分词后，采用向量模型进行存储，较好地保存了文献的结构及关键词在文中的顺序。

3 核心方法

下面将介绍中医文献语义关系图发现的详细方法。算法总体流程可分为文献关键词提取、高频关键词组计算、关键词组语义关系图识别 3 个步骤。图 1 描述了算法的总体流程。

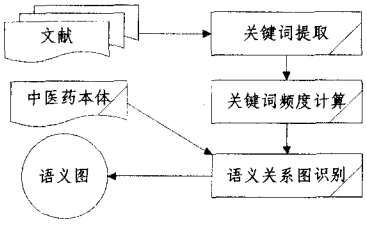

图 1 算法的总体流程图

算法的 3 个步骤简要描述如下：

1. 以本体概念名称为字典，对目标文献以句子为单位进行关键词的提取。关键词采用向量模型进行存储，并且保留在句子中的顺序。

2. 在关键词向量中寻找高频关键词组。采用关联规则算法的变异算法来计算各关键词组的频度，采用支持度作为阈值。

3. 利用本体知识库对高频关键词组进行语义关系识别；对不能识别的语义关系，利用知识库进行语义关系预测。最后生成一个语义关系图，关键词为顶点，连接顶点的有向边为其语义关系。

3.1 关键词提取

语义图发现算法的第一步是提取出文本中的关键词，这也是大多数文本处理、文本挖掘算法的第一步，也是非常重要的一步，对文本进行准确的分词，是后续算法良好运行的基础。

当前的自动化中文分词算法大致可以分为 3 类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。基于字符串匹配的分词方法，也称为基于字典的分词方法，其优点是分词准确性较高，其不足是需要一个全面的字典库，另外对新词的发现能力不足。基于理解的分词方法利用句子的句法和语义进行分词，这种方法需要大量的语言知识，汉语的语法非常复杂，词语数量庞大，因此这种方法还有很长的路走。基于统计的方法不需要字典库和语法知识，主要思想是文本中多次同时出现的字有一定的概率是词语，这种方法分词速度快，无需庞大的知识库，可以识别出一些新词，但是目前识别率较低，还有待研究。

本文中的语义图关系发现算法，需要将关键词利用本体在知识库中进行识别和预测，识别需要在知识库中存在描述，预测时也需要知道关键词的上位词，因此我们处理的关键词需要在知识库中已经定义。中医药文献特别是中医药古文献的自动化分析处理目前还存在许多困难。中医药文献所具有的一些特点，比如中医药学的概念表达、术语使用甚至语句内容都具有领域性，使得基于语法的自然语言处理方法在分析处理中医药文献中效果并不理想。基于上述原因，我们的分词算法采用了基于字典匹配的分词方法。

语义图发现算法处理的关键词需要在知识库中定义，因此我们将知识库中各本体的概念的名称作为关键词提取算法的字典。这样关键词字典的词属于中医药的范畴，因而提取到的关键词也属于中医药范畴，为后续算法提供了比较纯的数据源，提高了语义关系图的质量，减少了无关信息。

在算法的第三步语义识别和预测阶段，关键词数量的增长将极大地增加算法的时间和空间复杂度。因此关键词组内关键词的数量不宜过大。句子是一个相对较小而又有完整语义的单元。句子通常描述一个完整的语义，而句子内部的词语有比较密切的关系，表达某种语义关系，比如修饰名词的形容词和被修饰的名词之间的关系，谓语两边的名词之间的关系。因此关键词提取应以句子为单位进行。

向量模型是文档处理中常用的一种存储结构，其多维性和有序性的特点，可以较好地保存词语在文档中的结构位置。本文中关键词识别结果采用向量模型进行存储，并且保持其在原来句子中的前后位置关系。每个句子用一个向量进行存储。见式(1)，关键词向量 V 由 m 个关键词构成。

$$V=\{W_1,W_2,\cdots,W_m\} \tag{1}$$

向量中的关键词是有顺序的，按照关键词在句子中出现的顺序来排列，保持了关键词在文本中的前后位置。关键词之间的位置关系，可以作为关键词之间是否存在语义关系及

其语义关系重要程度的一种依据。

本文中待处理的文本称为数据源。数据源可以是几篇相关的文献，也可以是一篇文献，或者是文献中的一段文字。数据源最好是一些主题相同的文献，介绍同一类知识的文章，以便更好地计算高频关键词组，这将在第二步的算法中提到。

关键词提取的具体步骤如图 2 所示。

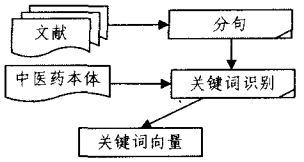

图 2　分词算法的流程图

第一步将数据源文献以句子为单位进行分解。第二步以中医药本体概念为字典，对每个句子进行中医药关键词的提取。关键词提取的结果以向量进行存储。

式(2)表示数据源进行关键词提取后的结果，集合 K 是关键词向量 V 的集合。

$$K=\{V_1,V_2,\cdots,V_n\} \tag{2}$$

集合 K 中包含了所有从数据源中提取出来的关键词，后续算法将在其中寻找潜在有价值的语义关系图。

此处所说的有价值的语义关系图的价值是指文献中描述的语义关系，反映了只有文献语义的语义图才能发挥文献的价值。

3.2　高频词的计算

关键词提取完成后，将进行高频关键词组的发现。多次一起出现的词组，其内部存在一定的有价值的语义关系的可能性较大，这是显而易见的。这也是关联规则数据挖掘算法的思想之一，本文高频词计算方法也借鉴了关联规则算法。

出现次数越多的关键词组，其越有可能存在有价值的语义关系(有价值是指文献中论述的语义关系)。据此对各种数量的关键词组进行出现次数的统计。关键词组的出现次数的阈值设为 E，当出现次数达到 E 次，那么该关键词组即为高频关键词组。这个 E 也就是关联规则查找算法中的支持度。这里不存在一些词语比如副词、连词、介词等多次出现的干扰问题，因为分词程序分词后得到的关键词都是中医药范畴的词语。

高频关键词组内的关键词需要保持在关键词向量中的顺序，不过可以在关键词向量中隔开 m 个词出现。关键词之间的位置关系隐含着词语之间是否存在有价值语义关系的依据，如果两个词的位置相近，那么两者间存在有价值的语义关系的可能性就较大；反之，如果两个词的位置相距较远，那么其两者之间存在有价值的语义关系的可能性就较小。关键词不要求连续出现，可以相隔 k 个词出现。因为相同语义的关键词组，可能会以不同的形式呈现，比如："大黄，泻火，凉血"和"大黄，泻火，大黄，凉血"虽然是同一种语义，形成相同的语义图，但是其关键词的数量是不同的，如果在计算关键词出现的阈值时，关键词可以隔开 1 个词的话，那么"大黄，泻火，凉血"就可以算出现了两次。

在一定间隔范围内，关键词对前后位置关系不敏感，也就是说，$\langle W_i,W_j\rangle$ 和 $\langle W_j,W_i\rangle$ 有等价的语义作用。比如"大黄治疗咽肿"和"咽肿被大黄治疗"在现实世界中的意义是相同的，而且预测算法将会对两种位置关系的关键词都进行语义关系识别。在一个完备的知识库中，对这两种关系都会进行描述，这样的话，无论 $\langle W_i,W_j\rangle$ 还是 $\langle W_j,W_i\rangle$ 都可以在知识库中获得识别。如果知识库中对这样的语义关系只描述一种，那么在语义识别算法中，对关键词进行正向和反向的识别，即在识别 $\langle W_i,W_j\rangle$ 的同时，也对 $\langle W_j,W_i\rangle$ 进行识别。如果两个关键词的间隔比较远，本文认为其两者之间不存在有价值的语义关系，因此要在一定范围内讨论关键词的前后位置关系。

下面将对高频关键词计算进行详细的描述。对于关键词向量集 K 的 V_i，设有 n 个关键词，分别计算 V_i 的各种数量的关键词组在 K 中的出现次数，如果达到阈值 E，则对应的关键词组属于高频关键词组。对两个以上的关键词进行出现次数的计算，从最大的关键词数量开始计算，比如 V_i 有 n 个关键词，那么从 n 个关键词开始计算，如果 n 个关键词属于高频关键词，那么小于 n 数量的关键词就不必再进行统计，因为它们都属于高频词。高频关键词组中的词，不必前后相邻，可以相隔 m 个关键词出现。

定义高频关键词集合 FK，对关键词向量集合 K 中的每个向量 V_i 进行如下步骤的计算：

Step 1　定义变量 n，n 为向量 V_i 中的关键词个数。

Step 2　在向量 V_i 中，对每组数量 n 的关键词组，不要求连续，可以相隔 k 个词，首先判断集合 FK 中是否存在该关键词组。如果存在，或者包含该关键词组的关键词组存在，则不必再进行计算。如果不存在，则进行计算，在其他关键词向量中寻找该关键词组，如果该关键词组出现数量达到阈值 E，则将该关键词组加入集合 FK 中。

Step 3　将 n 减 1。如果 n 不小于 2 则转到 Step2；否则算法完成，集合 FK 中的便是发现的高频关键词组。

3.3　语义关系图识别和预测

接下来是算法最核心的一个步骤，对高频关键词组进行的语义关系图的识别，包括识别和预测语义关系两个步骤，最终将语义关系连成一个语义关系图。

3.3.1　语义本体知识库

语义本体知识库中，本体包含了名称、定义等属性，描述了本体的基本信息。而三元组〈本体 a，语义关系，本体 b〉描述了本体 a 和本体 b 之间的语义关系。在此基础上整个知识库中的知识相互关联，形成一个语义网络。

对于关键词之间的语义关系，如果已经在知识库中存在描述，则通过在知识库中查找可以确定该关系；如果在知识库中还没有描述，那么可以通过知识库中已有的相关语义关系对该语义关系类型进行预测。

3.3.2　语义关系识别

如果两个关键词之间的关系在知识库中已经存在描述，那么其语义关系可以通过查找直接确定。比如在知识库中存在〈大黄，功效，清热〉的语义关系，则对于关键词"大黄"和"清热"，可以直接确定大黄与清热之间是功效的语义关系，即大黄有清热的功效，因此大黄和清热之间的语义关系得以识别。

在中医药本体知识库中，有一种语义关系是"正名关系"，即概念的正规名称，相对于正名的是异名，正名和异名是相同事物的不同名称。比如：大黄是正名，其异名包括黄良、火参锦纹等。在中医药中，存在着很多的正名异名现象[10]，正名在语义本体知识库中有更全面的语义关系的描述。将属于异名的中医药概念词转换为正名，然后再进行语义关系的识别，

这将提高语义关系识别的成功率。

在中医药本体知识库中，还存在一种“上位词”的语义关系（见图 3），即概念之间的上下级的关系，或者说是父子关系。在此知识库中，父概念之间所拥有的关系，子概念是可以继承的。比如大黄属于阴性药物，与症候上火具有治疗的语义关系。概念的上位词与上位词之间可能也存在语义关系，但是越上层的语义关系，其抽象性越大，对现实的指导意义却越小。

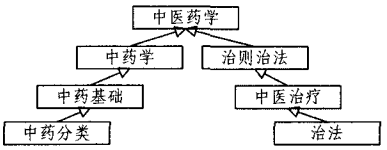

图 3　中医药上位词关系举例

利用上述这些特点制定对应的语义关系识别算法。首先直接查找关键词之间的语义关系。如果语义关系查找失败，那么查找概念的逐个层次的上位词之间的语义关系。离关键词越近的上位词，其与其他概念的语义关系越具有指导意义。因此本文将选取离概念词最近的 t 层以内的上位词之间的语义关系作为关键词之间的语义关系。

接下来介绍语义关系识别的具体步骤。设关键词向量 V 为包含了 n 个关键词的向量，W_i 为关键词，见式(3)。

$$V=\{W_1,W_2,\cdots,W_n\} \tag{3}$$

对于向量中的关键词，$W_i\in V$，$W_j\in V$，$i\neq j$，语义关系识别算法如下：

Step 1　在语义本体知识库中，查找以 W_i 和 W_j 为本体名称的概念词。如果找到了，算法就此结束。

Step 2　检查 W_i 和 W_j 是否是异名，如果是异名，找到其对应的正名。然后用其正名来进行语义关系的识别，如果知识库中存在其语义关系描述，那么将该语义关系作为 W_i 和 W_j 之间的语义关系，算法就此结束。

Step 3　找到 W_i 和 W_j 的上位词，包以及其上位词的上位词，包括追溯到最上层的上位词的全部上位词。这样得到了 W_i 和 W_j 的上位词集合，在集合中，保留上位词的层级关系。首先在概念词的直接上位词之间进行语义关系的查找，如果没有找到语义关系，那么向上层的上位词追溯，结合上层的上位词进行查找。不断地查找，不断地加入更上层的上位词，直到找到语义关系，或者到达最上层的上位词为止。

为了得到一个信息丰富的语义关系图，本文对关键词向量 V 内的所有关键词进行语义关系的识别。即将 V 中包含的 n 个关键词进行两两配对（不与自身配对），n 个关键词，将产生 $n*(n-1)/2$ 个词对。然后对这些词对在本体知识库进行语义关系的查找，如果知识库中存在描述，那么识别成功，记录下其语义关系。

3.3.3　语义关系预测

对于没有找到对应语义关系的词对，将进行语义关系预测。位置间隔越小的关键词对，其存在有价值的语义关系的可能性越大。这是因为出现在一个句子中的词语，其位置越近，句子阐述其语义关系的可能性越大。设位置间隔小于 m 的关键词之间存在语义关系的可能性较大，我们将对位置小于 m 的关键词对的语义关系进行预测。

经过分析，得到定理 1 如下。

定理 1　在知识库中，设 $R=\{r_i\}$ 为本体 A 和本体 B 各自所有的下位词（其子概念，相对于上位词）之间的语义关系的集合。当本体 A 和本体 B 的下位词及这些下位词之间的语义关系足够完备时，对于本体 a 和 b，a 属于本体 A 的下位词，b 属于本体 B 的下位词，那么 a 与 b 的下位词之间的语义关系 r_i 近似服从概率 $P(r_i)$。

$$P(r_i)=\frac{\sum r_i}{\sum_{i=0}^{n} r_i} \tag{4}$$

简单描述一下该定理的证明。如果本体 A 和本体 B 的所有下位词的语义关系都在知识库中有描述，随机地从本体 A 的下位词中选择一个词 a，同样随机地从 B 的下位词中选择一个下位词 b，那么 a 和 b 之间的语义关系的可能性是服从概率 $P(r_i)$ 的。在现实情况下，知识库不可能完全包含所有下位词之间的语义关系，所以只能说知识库越完备，语义关系 r_i 就越接近概率 $P(r_i)$。

在知识库中，词 a 和词 b 之间的语义关系有一定的概率是不存在的。我们的预测算法中，排除了这一情况，因为我们计算的两个关键词是高频词，假定其存在一定的语义关系。

基于以上的分析，我们设计语义关系预测算法。预测算法首先检查两个词的步长是否在阈值 m 内，然后检查是否已经识别到了语义关系。如果是符合条件的关键词对，先找到这两个关键词的上位词，计算上位词的子概念之间的各种语义关系的数量，然后在此基础上计算各种语义关系的概率。如果两个关键词同属于一个上位词，也是适用于定理 1 的，在预测时，按照不同上位词的情况处理即可。

设关键词对 a 和 b 的上位词分别是本体 A 和 B，其位置间隔数小于 m，并且其之间的语义关系在本体知识库中识别失败。另设 $FS(a)$ 为关键词 a 的上位词的集合；$CS(s)$ 表示 s 的下位词的集合。下面介绍语义关系预测算法的具体步骤。

Step1　找到关键词 a 和 b 各自的上位词的集合，即 $FS(a)$ 和 $FS(b)$。

Step2　统计上位词的所有下位词之间的各种语义关系的数量。设 $wordSet(a)=\{k|k\in CS(s),s\in FS(a)\}$，$wordSet(b)=\{k|k\in CS(s),s\in FS(b)\}$。将 $wordSet(a)$ 和 $wordSet(b)$ 进行笛卡尔积的配对，然后将笛卡尔积的各关键词对进行语义关系查找，并且统计各种语义关系的数量。

Step3　根据式(4)计算各种语义关系的概率。语义关系 $r=r$ 的数量/其他语义关系的数量。

图 4 是语义关系预测算法模型图，小写字母是待预测的关键词，大写字母是其对应的上位词，虚线表示语义关系。

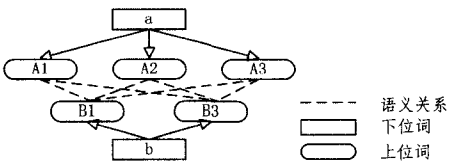

图 4　语义关系预测模型图

3.3.4　语义关系图

当语义预测过程结束后，根据前面的计算结果，将关键词组中的词语用语义关系进行连接，形成语义关系图。该语义关系基于语义识别和语义预测相结合的语义关系图。图 5 是 3 个关键词可能得到的一种结果。

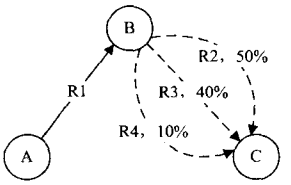

图 5　语义关系图发现结果示例图

A 和 B 之间的语义关系在知识库中已经描述，用实线表示它们之间的语义关系；B 和 C 之间的语义关系在知识库中未直接描述，因此采用预测算法对语义关系进行了预测，预测到的语义关系用虚线表示，虚线上标明了语义关系的名称，还有语义关系存在的可能性，百分数越大，语义关系存在的可能性越高。

语义关系图可以帮助相关研究人员快速查看中医药文献中包含的语义关系，预测的语义关系可以帮助研究人员发现新的语义知识。

4　算法实验

4.1　实验环境

实验以中医药语义本体(TCM Semantic Ontology)为知识库，知识库中拥有 30 万本体概念，并且对 40 余万本体之间的语义关系进行了标注[6]。另外，知识库中定义了 56 种语义关系类型，这些语义关系包括上位词、下位词、治疗、引起等。中医药语义本体知识库，以 RDF 为元数据进行本体的表述，用关系数据库进行数据存储。我们将这 30 万本体概念的名称作为关键词字典，对目标文献进行分词。知识库中的 40 万的语义关系描述将用于语义关系图的识别。

实验程序用 Java 语义实现。Java 可用的开源组件比较多，而且良好支持跨平台运行。实验分词采用 MMSeg4j 分词程序。MMSeg4j 是用 Chih-Hao Tsai 的 MMSeg 算法[5]实现的中文分词程序。MMSeg4j 能很好地支持词库的扩展，甚至可以将原有词库进行替换。本实验中，将 MMSeg4j 的原有词库替换成中医药本体概念的名称组成的词库。

程序对本体的语义关系图发现方法进行了一体化的设计，即将关键词的提取、高频关键词的统计、语义关系图的生成设计成一个完成的程序。程序读入数据源以后，直接可以输出相应的语义关系图，提高了效率，使用也比较方便。实验结果中的语义关系图以规范的格式输出，可以采用一些程序组件，将文本格式的结果以图形化的方式更生动直观地显示。

4.2　实验过程及结果

实验对多种不同类型的中医药文献进行了测试，包括中医药的一些古文献和一些现代的文献，按文献的内容又可以分为病案类、药物类、治法类等。

在程序读取前，将待处理的文献转化成文本文件，以便程序读取。高频关键词识别时，关键词的间隔 k 设为 1，即高频关键词组内的词可以相隔一个词出现；预测关键词的距离 m 设为 1，即将相隔一个词以内的关键词对进行语义关系预测。

对实验结果进行检查后，证实该方法可以对中医药文献进行语义关系图发现。下面对实验中的两组结果进行分析。

实验结果 1　高频关键词组是“大黄，蓼科，中国，马蹄大黄”，对其进行语义关系识别和预测后得到结果如表 1、表 2 所列。

表 1 是对关键词组进行语义识别后得到的结果，表示已经在知识库中存在描述的语义关系。可以看到大黄和蓼科之间是下位词关系，马蹄大黄和大黄是异名关系，马蹄大黄生产于中国。表 2 是预测算法根据知识库中的已有知识进行的预测。蓼科和中国之间没有预测到对应的语义关系，这是符合我们预期的。大黄和马蹄大黄之间是异名关系，即相同概念的关系，所以我们没有再对大黄和中国、蓼科和马蹄大黄进行语义关系预测，因为马蹄大黄和中国、蓼科和大黄的语义关系知识库中已经存在描述了。

表 1　语义关系识别结果

概念名称	概念名称	语义关系
大黄	蓼科	下位词
马蹄大黄	大黄	异名
马蹄大黄	中国	生产于

表 2　语义关系预测结果

概念名称	概念名称	语义关系
蓼科	中国	无

图 6 是该组关键词的语义关系图。图中带箭头的实线表示识别到的语义关系，箭头的方向是语义关系的方向。

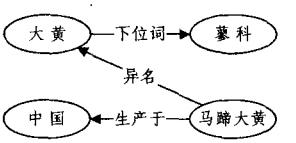

图 6　语义关系图

实验结果 2　高频关键词组“大黄，清热，解毒，咽肿”，此处将 m 值设为 3，在进行语义关系识别和预测后得到如表 3 所列的结果。

表 3　语义关系识别结果

概念名称	概念名称	语义关系
大黄	清热	功效

从表 3 可以看到，这组高频关键词中，大黄和清热之间存在着功效的关系，在知识库中已经存在描述。

下面来看预测的情况，表 4 中大黄的上位词是泻火药、攻下药、蓼科，而解毒的上位词是病症防治、泻火药、攻下药，蓼科和病症防治的下位词之间的语义关系 52%的概率是功效，还有 37%是预防，这是符合事实的。清热的上位词是寒热治法，其与病症的下位词之间有 31%的概率是同时发生的关系，概念上相关的概率是 41%，还有 20%的概率是先后发生。

表 4　语义关系预测结果

概念名称	概念名称	语义关系
大黄	解毒	功效，52% 预防，37% 其他，11%
清热	解毒	同时发生，31% 概念上相关，41% 先后发生，20% 其他，7%
清热	咽肿	…的方法，65% 其他，34%
大黄	咽肿	治疗，76% 其他，24%
解毒	咽肿	无

(下转第 251 页)

改进后的GSO算法用在多峰函数问题中，仿真实验结果表明，改进后的GSO算法具有在种群规模较小、迭代次数较少的情况下也可以精确捕获函数定义域内所有峰的特点，同时消除了人工萤火虫群漫无目标的随机移动的局限。本算法不仅保持了基本GSO算法的特点，而且提高了算法的运行效率。由于GSO理论和应用研究还处于不断完善阶段，因此许多问题有待于人们不断地探索和解决。比如，GSO对于其参数的选择有较高的敏感度，参数选择的好坏直接影响结果的精度。因此，如何选择恰当的参数和对GSO算法进行收敛性分析，以及与其它智能算法有机地融合，设计出一些高效的群混合智能算法并用之求解实际问题，将是我们今后所要做的研究工作。

参 考 文 献

[1] Wei L Y, Zhao M. A niche hybrid genetic algorithm for global optimization of continuous multimodal functions[J]. Applied Mathematics and Computation, 2005, 160(3): 649-661

[2] 李敏强，寇纪淞. 多模态函数优化的协同多群体遗传算法[J]. 自动化学报，2002，28(4)：497-504

[3] Krishnanand K N, Amruth P, Guruprasad M H, et al. Glowworm-inspired robot swarm for simultaneous taxis towards multiple radiation sources[C]// IEEE International Conference on Robotics and Automation. Orlando, Florida, May 2006: 958-963

[4] Krishnanand K N, Ghose D. Detection of multiple source locations using a glowworm metaphor with applications to collective robotics[C]// Swarm Intelligence Symposium. June 2005: 84-91

[5] Krishnanand K N, Ghose D. Theoretical foundations for multiple rendezvous of glowworm-inspired mobile agents with variable local-decision domains[C]// American Control Conference. June 2006: 14-16

[6] Krishnanand K N, Ghose D. Theoretical foundations for rendezvous of glowworm-inspired agent swarms at multiple locations [J]. Robotics and Autonomous Systems, 2008, 56(7): 549-569

[7] 李晓磊，邵之江，钱积新. 一种基于动物自治体的寻优模式：鱼群算法[J]. 系统工程理论与实践，2002，22(11)：32-38

[8] Krishnanand K N, Ghose D. Glowworm swarm optimisation: a new method for optimising multi-modal functions[J]. Int. J. Computational Intelligence Studies, 2009, 1(1): 93-119

[9] Brits R, Engelbrecht A P, van den Bergh F. A niching particle swarm optimizer[C]// The 4th Asia-Pacific Conference on Simulated Evolution and Learning. 2002: 692-696

[10] Parsopoulos K, Vrahatis M N. On the computation of all global minimizers through particle swarm optimization[J]. IEEE Transactions on Evolutionary Computation, 2004, 8(3): 211-224

[11] Fevrier V, Patricia M. Parallel evolutionary computing using a cluster for mathematical function optimization[C]// The Annual Meeting of the North American Fuzzy Information Processing Society. 2007: 598-603

[12] Muller S D, Marchetto J, Koumoutsakos S A P. Optimization based on bacterial chemotaxis[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(6): 16-29

(上接第217页)

最终语义关系图如图7所示，带箭头的虚线表示预测的语义关系，百分数表示该语义关系存在的可能性。

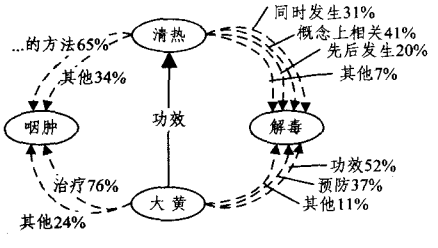

图7 语义关系图

结束语 本文提出了一种利用中医药本体知识库对中医药文献进行语义关系图发现的方法。本文的关键词提取采用中医药本体的概念名称作为字典，关键词提取无需依赖语法分析程序。分词后采用类似关联规则的算法进行高频关键词计算。然后利用中医药本体知识库对高频关键词组进行语义关系图的识别和预测。最后本文通过实验对提出的算法进行了验证，结果证实了算法的可行性。

利用语义关系图对中医药文献进行知识发现，具有表达直观性，知识获取全面性及可推理性等优点，能够有效帮助相关研究人员对大量文献进行知识获取、分类、推理等。

本文提出的语义关系图发现方法，并不仅限于中医药领域，理论上说如果有对应的领域本体知识库，就可以对该领域文献进行知识发现。

本文提出的语义关系图的发现方法是一种初步的探索，还有很多方面有待继续探索。比如在关键词的提取部分，结合基于统计的方法和一些额外的词库对新词进行发现；在语义关系预测部分，结合语法进行语义关系类型的预测。

参 考 文 献

[1] Berners-Lee T, Hendler J, Lassila O. The Semantic Web[J]. Scientific American, 2001

[2] Shadbolt N, Berners-Lee T, Hall W. The Semantic Web Revised [J]. IEEE Intelligent Systems, 2006, 21(3): 96-101

[3] Takaaki H, Sekine S, Grishman R. Discovering relations among named entities from large corpora [C]// Proceeding of Conference ACL 2004. Barcelona. Spain: Association for Computational Linguistics, 2004: 415-422

[4] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases[C]// SIGMOD Conference. 1993: 207-216

[5] Tsai C H. MMSEG: A word identification system for Mandarin Chinese text based on two variants of the maximum matching algorithm[R]. 2000

[6] Chen H J, Wu Z H. Semantic Web Model, Methodology and Applications[M]. Springer-Verlag, GmbH, 2008

[7] Salton G, Wong A. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613-620

[8] Leskovec J, Grobelnik M, Milic-Frayling N. Learning sub-structures of document semantic graphs for document summarization [C]// KDD 2004 Workshop on Link Analysis and Group Detection(LinkKDD). Seattle, Washington

[9] Zhang Xiaogang, Chen Huajun. Ontology Based Semantic Relation Verification for TCM Semantic Grid[C]// ChinaGrid Annual Conference. ChinaGrid '09. Fourth, 2009: 185-191

[10] 何前锋，尹爱宁，刘静，等. 中医药同异名现象与标准研究[J]. 中国中医药信息杂志，2008(S1)