DOI:10.13374/j.issn1001-053x.2007.s2.104
第29卷 增刊2 北京科技大学学报 Vol.29 Suppl.2
2007年 12月 Journal of University of Science and Technology Beijing Dec. 2007

基于图分析的领域知识获取技术

张德政 阿孜古丽 刘洁卉

北京科技大学信息工程学院, 北京 100083

摘 要 知识获取技术制约着知识系统的研究和应用,有效地从文本中提取领域知识成为知识获取的重要途径. 本文提出了基于本体和图分析的领域知识获取技术,分析了本体数据结构、本体概念的实例化以及基于图分析的语义场构造方法. 建立了具有广泛适用性的文本知识获取系统框架,实现了原型系统. 通过针对中医医案的中医领域知识获取实验验证,取得了较好的效果.

关键词 本体;图分析;知识获取

分类号 TP182

知识获取是专家系统乃至其他知识系统的亟待解决的问题. 在关联规则数据挖掘研究基础上,数据挖掘逐渐拓展到Web挖掘、生物信息等包含大量复杂类型数据的众多实际应用领域[1-2]. 数据挖掘与知识发现部分地解决了知识获取问题,对于多关系并蕴含复杂结构的知识的获取迄今没有有效的方法. 充分利用已有知识,来获取知识结构中的那些未知关系,或利用领域常识来获取更具针对性和概括性的知识,已经成为知识获取、知识发现研究的重点[3]. 本文依据认知机理,从领域基础知识出发,通过文本处理,建立概念之间的关系,进而通过分析概念构成的图结构特征来获取领域专家感兴趣的知识.

1 知识获取的相关工作

对于多关系以及结构知识获取的研究工作主要包括两类:一类是针对挖掘特定的知识结构,获取知识结构的频繁子图,其工作主要集中在基于图的算法研究上,如AGM[4]和FSG[5]等子图的挖掘算法,算法都利用邻接矩阵分别对图的顶点和边进行逐层构造,以最终获取频繁的子图. 另一类是从图中寻找某种有用的属性信息,如度、最短路、路径中包含的信息量等,来凸现图节点间的关系或差异. 通过充分地反映出节点在图中的位置特性,将图中节点的显著性进行"放大"来定义节点的重要性;度量方法主要包括节点的度(Degree),亲近度(Closeness),中介性(Betweenness),信息(Information),特征向量(Eigenvector)等[6].

这两类方法分别从传统意义下的数据挖掘角度和图属性分析的角度考虑了知识获取. 这些方法在各自的适应范围内取得了较好的效果. 但对于深层次知识、隐性知识以及小样本数据情况下知识获取受到局限,同时目前也未见到其他有效方法. 同时,在实际应用中图结构的构造制约着相关技术的应用.

挖掘作为从文本中获取知识的技术之一已经有着深入的研究. 文本挖掘通过分类聚类、自动文摘等技术从大量文档中抽取有用的结构与模式,所获取的知识为层次较高的抽象知识[7]. 中医医案是一类特殊的文档,名老中医学术思想与临证经验是通过传承、实践以及创新而形成的独特知识体系,知识隐含在名老中医辨证施治过程以及所形成的医案之中. 有效地获取医案中的知识,深层次挖掘隐藏在诊疗过程中的隐性知识,最大限度地获取与保留名医数十年积累的诊疗经验,是实现中医传承亟待解决的关键问题之一. 由于疾病机理的复杂性和中医诊疗的个性化特征,相似或相近疾病或病证的数量较少,对于小样本医案采用常规文本挖掘技术来获取深层次知识的难度较大.

基于应用与理论研究需求驱动,将领域知识与图结构分析融合在一起来获取文本中的知识,并研发了相应的软件系统KAS(Knowledge Acquisition System). 系统构建了中医领域本体知识库,实现中医文本处理相关技术,进而基于图结构分析来完成个体知识的实例化,获取名老中医诊疗个性化的独特的知识.

收稿日期:2007-10-12

基金项目:国家"十一五"科技支撑计划基金资助项目(No.2007BAI10B06);国家"863"计划基金资助项目(No.2003AA115220)

作者简介:张德政(1964—),男,副教授,博士

2　结构知识及知识获取

复杂知识体系显化于知识结构之中就构成了一类结构知识. 结构知识表现为知识节点以及知识节点之间的关系, 其中知识节点是由领域知识体系中的概念构成的, 概念之间的连线, 对应于概念节点之间的语义联系. 为了获取特定文本中的知识, 需要首先将文本中的概念及其之间的关系组织成相应的数据结构, 这里引入图论的方法[8]. 如果将知识结构中的概念定义为图的顶点, 概念之间的联系定义为图的边, 知识结构可用有向图结构来表示. 一个有向图 $\boldsymbol{D}$ 是一个三元组(V, E, f), 其中 V 是一个非空的集合, 它的元素称为有向图 $\boldsymbol{D}$ 的结点, E 是一个集合, 它的元素称为有向图 $\boldsymbol{D}$ 的弧(边), f 是一个从E 到 $V\times V$ 上的映射(函数), 例如, $V=\{a, b, c, d\}$, $E=\{e_1, e_2, e_3, e_4, e_5, e_6\}$, 且 $f(e_1)=<a, b>$, $f(e_2)=<c, b>$, $f(e_3)=<b, c>$, $f(e_4)=<c, d>$, $f(e_5)=<d, b>$, 则 $\boldsymbol{D}=(V, E, f)$是一个有向图, 如图 1 所示.

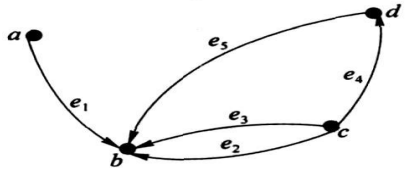

图 1　知识结构的有向图示例

个性化的独特的知识可用通过演绎推理来获取, 即知识结构中所包含的独特的、规律性的知识. 在知识结构中, 知识是通过知识结构的图结构特征来表示的. 常见的图结构特征有: 度、核数的度指标以及接近度等指标[9].

(1) 度(degree). 度代表图中某一个节点与之相连的边的数目. 在知识结构中表达一类语义联系, 用无方向图来定义.

(2) 子图(subgroup). 一个图 $\boldsymbol{G}$ 的子图 $\boldsymbol{G}_s$ 的定义为: $\boldsymbol{G}_s$ 的点集是 $\boldsymbol{G}$ 的点集的一个子集, 并且 $\boldsymbol{G}_s$ 中的线集 $\boldsymbol{L}_s$ 也是 $\boldsymbol{G}$ 的线集的一个子集, $\boldsymbol{G}_s$ 中的所有线也必须是 $\boldsymbol{G}$ 中所有点之间的线. 由于 $\boldsymbol{L}_s$ 是 $\boldsymbol{L}$ 的一个子集, 图 $\boldsymbol{G}$ 中的两点可能包含在子图中, 但是, 图 $\boldsymbol{G}$ 中的连接这两个点之间的线不一定在子图 $\boldsymbol{G}_s$ 中.

(3) 核数(cores). 设图 $\boldsymbol{G}=(\boldsymbol{V}, \boldsymbol{E})$ 是一个图. $\boldsymbol{V}$ 是顶点集合, 并且 $\boldsymbol{L}$ 是线集合(边或弧). 用 n 表示顶点的个数 $n=|\boldsymbol{V}|$ 并且用 m 表示边的个数 $m=|\boldsymbol{L}|$. 由集合 W 产生的子图 $H=(W, L|W)$ 是 k 一核或者说是次序为 k 的核, 当且仅当对任意 $v\in W: \deg(v)\geqslant k$ 并且 H 是具有这个特性的最大子图. 顶点 v 的核数是包含这个顶点的核的最高次序.

知识获取过程都是以原有的知识为基础来获取新的知识. 用一种清晰的方法把领域知识分解为一组知识元以及它们之间的相互关系, 这些知识元和相互关系组织在一起就构成了领域本体, 即生成一个领域知识库. 在中医领域, 通过对领域概念和概念关系的组织成本体, 即生成了中医基础知识库. 利用领域知识通过演绎推理可以获取个性化的知识. 为了把本体知识库存储起来, 组织了邻接链表作为图的存储结构, 便于图的遍历和查找. 在这些资源的基础上, 可以建立知识获取的系统框架. 对文档进行分析处理, 生成了新的知识, 达到了知识获取的目的[10].

3　系统结构与功能

利用上述理论和相关技术来构造基于本体的领域知识获取系统——KAS. 系统主要包括文本预处理模块、知识库模块和知识获取等模块, 系统结构如图 2 所示. 文本预处理模块负责对文本进行词法分析, 提取出文本的特征词. 知识库模块负责中医领域本体的组织、数据结构表示. 知识获取模块负责文本特征词的实例化, 组织形成语义场, 完成获取知识.

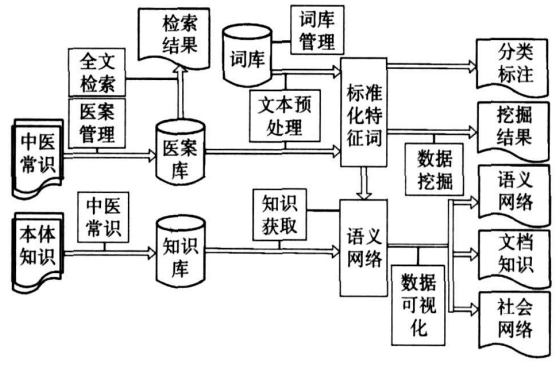

图 2　知识获取系统结构图

3.1　文本预处理

中文分词是基础工作, 系统采用基于统计模型的最大概率法进行分词. 最大概率法相当于一元语法模型, 每次选择出现概率最大的路径作为切分结果. 运用一元语法模型可以达到 90%以上的切分正确率. 利用大规模的中医语料库和成熟的 n 元语法统计模型, 可将切分正确率提高很多.

在对文本进行分词处理后, 系统采用了预制停

用词库的方法, 可以达到滤词和抽取特征词的作用. 首先对大量病例病案进行切分和标注, 训练出词频和词性的信息, 把词性为助词、代词、介词、语气词、语素词作为停用词的候选词汇. 从候选停用词中, 选择那些可能会在文本中频繁使用, 而无宜于语义表达的词语作为停用词. 本文最终构造停用词 800 个, 包括符号 40 个. 在对文本进行滤词后, 可以抽取出特征词, 结合本体概念词库和本体概念同义词库, 抽取出文档对应本体网络的标准概念.

3.2 知识库

知识结构与本体及其关系有着很好的对应性. 本体中的关系表示概念之间、概念和个体实例之间的关联. 领域本体是用于描述指定领域知识的本体, 它由概念、关系和子领域本体组成. 开发一个本体的过程包括定义本体中的类、定义概念之间的关系. 通过添加特定的属性插件赋值信息和限制条件, 就可以建立起一个知识库. 本文构建了中医领域本体知识库, 并将部分中医领域本体中药症关系采用可视化的形式表示出来, 如图 3 所示.

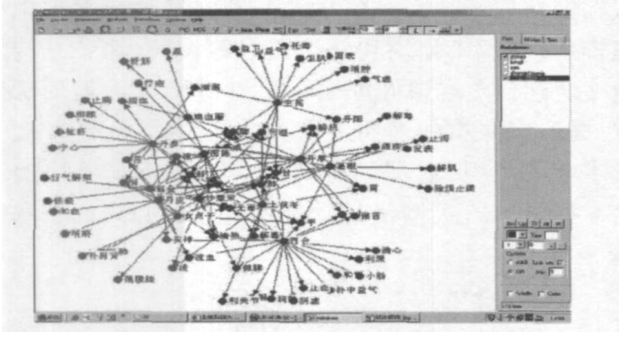

图 3 知识库本体关系示意图

基于本体结构以及图与数据结构的对应关系, 本文构建了中医领域本体的网状数据结构[11], 定义了结点类型和邻接表内条目类型如下:

```
Struct Vertex //顶点基本信息
{
  string name; //本体概念结点
  vector<Edge> adj; //邻接结点
  Vertex *prev; //前一个结点
  int scratch; //相关信息
};
//邻接表内条目
struct Edge
{
  Vertex *dest; //第二个本体概念结点
  string relation; //概念间的结点关系
};
```

对知识库中的概念及关系组织成网状数据结构, 以各个概念名称作为头结点, 与它相关的概念和概念之间的关系作为邻接链表的结点. 概念抽取后得到的是相关的特征词的集合, 为了对应到本体的概念, 需要对特征词进行本体术语概念的标准化. 这里采用统一词库的方法, 把标准术语概念和特征词之间作映射. 一个本体概念可以对应多个特征词和不规范词汇.

4 基于图的知识获取

4.1 概念实例化

概念标准化后, 就可以对文本分析而来的概念进行实例化的映射. 把与文档相关的概念以及概念间的关系从本体知识库中抽取出来, 形成多个概念及概念关系的语对. 把这些语对再组织成网状的数据结构形式, 在此基础上就可以进行语义场的构建.

4.2 概念语义场

语义场指语义有关联的词共同构成的一个集或区, 场内每个成员的意义取决于与成分之间的相互制约关系. 语义场形式上是词语的集合, 实质上是概念凭借语义关系共同作用、相互关联的一个集合. 按照其语义关系, 可将语义场划分为聚合场和联想场. 为了完成知识获取, 本文主要构造聚合场, 即由相关概念之间类聚关系而形成的概念集合[12].

4.3 知识获取

将文档中抽取出的概念作为原子概念, 由原子概念与知识库中的概念相匹配, 并根据知识库中概念之间的二元关系建立原子概念的语义链. 根据两个概念之间的二元关系, 就可以建立资源的语义链. 利用语义链结构与知识库中概念语义网络模型相对应, 其中节点表示概念, 有向边表示类型化的语义链. 在语义链的基础上就可以构建语义场, 如对某一味药材构造语义场, 可以构造与药材归经、药物功效、药病、药证、味、性和药毒等多种关系的语义场. 利用语义场和图结构的性质就可以获取由文档概念语义场所隐含的知识, 如计算图结构中重要概念的核心性的算法. 对于一个给定图 $\boldsymbol{G}=(V, L)$, 连续删除度数小于 k 的顶点以及和这些顶点相关的边, 余下的图就是 k—核. 算法结构如下:

输入: 图 $\boldsymbol{G}=(V, L)$

输出: 每个顶点的核数表

compute the degrees of vertices;

order the set of vertices V in increasing order of their degrees;

for each $v \in V$ in the order do begin

```
    core[ v] := degree[ v] ;
    for each u∈Neighbors( v) do
      if degree[ u] > degree[ v] then begin
      degree[ v] := degree[ u] -1;
      reorder V accordingly
    end
  end;
```

5　实验验证

为了验证算法的有效性, 对我国中医肝病专家钱英教授诊断肝病医案进行了分析. 在所分析的病例中病人主要症状为: 脉沉细, 舌质淡, 苔白厚, 舌下静脉粗, 手末梢暗, 眠差. 钱英教授认为肝藏血, 主疏泄, 达阳气于四末, 慢性肝病患者, 常有痰、瘀阻于肝络, 出现手背末梢发暗. 舌下静脉曲张、增粗亦往往为肝络不通之表现. 人体为统一的整体, 有诸内必形诸外, 体内血液循环受阻亦必形之于外. 医案按语给出病机分析与治疗是痰湿致肝络不通, 湿郁阻络, 气阴两虚, 补气益阴, 化湿通络, 其中以气阴两虚为病机的主要方面. 本医案辨为气阴两虚、湿郁阻络, 用益气养阴、化湿通络之法治疗.

利用 KAS 系统对此病案进行分析, 如图 4 所示, 可以看出, 本医案遣方用药以凉性、温性药物受关注程度较高, 均在 0. 5 以上, 而微寒性药物关注度较低. 用药药味以甘、苦、咸、辛者关注程度较高, 关注程度最高者为甘味, 其次为苦味, 再次为咸味, 而

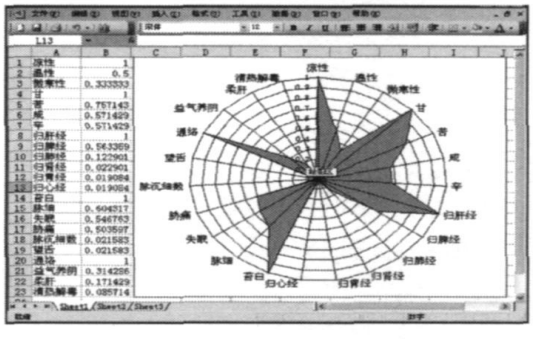

图 4　KAS系统肝病医案分析结果

其他药味的关注度较低. 甘味、苦味药物的关注程度均在 0. 7 以上, 最小值也在 0. 5 以上. 在医案中, 所用药物归经的受关注程度依次为肝经、脾经、肺经. 对医案症状分析中可以看出, 主要症状为苔白、脉细、失眠和肋痛, 其出现度均在 0. 5 以上. 在治法中, 以通络、益气养阴、柔肝为主要方法, 体现了中医的综合辨证治疗的方式.

6　结论

本文就基于本体的领域知识获取技术进行了探讨, 通过本体所表达的概念知识之间的关系与图结构分析的有机结合, 提供了一类文本知识获取技术. 结合中医医案知识获取, 构造了 KAS 原形系统, 通过实例检验验证了技术方法与系统的有效性.

参　考　文　献

[1] 周勇. 数据挖掘技术发展综述. 中国科技信息, 2005, 35(16): 35

[2] Cheng L H, Mu C C. Credit scoring with a data mining approach based on support vector machines. **Expert Systems with Applications**, 2007, (33): 847

[3] 陆汝矜. 人工智能. 北京: 科学出版社, 2000

[4] Inokuchi A. An apriori_based algorithm for mining frequent substructures from graph data //PDKK2000. Lyon France, 2000

[5] Kuramochi M. Frequent subgraph discovery // CDM 2001. San Jose, USA, 2001

[6] Wasserman S, Faust K. Social network analysis: methods and applications. Cambridge: Cambridge University Press, 1994

[7] 吕东煜, 党齐民. 基于文本挖掘的可视化竞争情报. 计算机应用与软件, 2005(2): 50

[8] 肖位枢. 图论及其算法. 北京: 航空工业出版社, 2005

[9] Wang Y, Xu J, Xi Y. The core and coritivity of a system. **Journal of Systems Engineering and Electronics**, 1993, 4(2): 1

[10] Zhu Z. Mining inter_ enttity semantic relation using improved transductive learing // Proceedings of the Second Internationl Joint Conference on Natural Language Processing (IJCNLP—05). Jeju Island, 2005

[11] Weiss M A. Data Structures and Problem Solving Using C++. 2nd. Pearson Education Inc, 2000

[12] 宋炜, 张铭. 语义网简明教程. 北京: 高等教育出版社, 2004

Domain knowledge acquisition on basis of graph analysis

ZHANG Dezheng, Aziguli, LIU Jiehui

Information Engineering of School, University of Science and Technology Beijing, Beijing 100083, China

ABSTRACT Knowledge acquisition technique gives constrains to study and application of knowledge system, therefore, effective extraction of domain knowledge from text becomes the major approach to knowledge acquisition. Domain knowledge acquisition technique based on theories of ontology and graph analysis is proposed, data structure of ontology, instantiation of ontology concept and construction of semantic field are discussed. Frame of text knowledge acquisition system is constructed, which is of general applicability, and the prototype system has been realized. By experimental verification of domain knowledge acquisition in Traditional Chinese Medicine, the system achieved good results.

KEY WORDS ontology; graph analysis; knowledge acquisition