●黄水清　熊　健　李志燕

闭合式非相关文献知识发现方法在中文文献中的验证

摘　要　在基于 Swanson 的闭合式与开放式知识发现法具体算法过程的基础上,以中国期刊网医药卫生大类的数据为测试集,采用的闭合式和知识发现方法,在中文科技文献中重现了"雷诺氏病和鱼油"、"偏头痛"和"镁缺乏"两对概念的知识发现过程,验证了 Swanson 的基于非相关文献的知识方法中的闭合式方法在中文文献中同样可行。表 2。参考文献 12。

关键词　知识发现　闭合式方法　Swanson　非相关文献
分类号　G254

ABSTRACT　On the basis of Swanson-based algorithm for closed and open knowledge discovery methods, with data from the "Medicine" class in CNKI as test set and with closed knowledge discovery method, the authors verify the applicability of Swanson's closed non-interactive literature knowledge discovery method in Chinese literature. 2 tabs. 12 refs.

KEY WORDS　Knowledge discovery.　Closed process.　Swanson.　Noninteractive literature.
CLASS NUMBER　G254

1986 年,Don R. Swanson 提出了非相关文献知识发现方法。这种方法可以揭示蕴涵于公开发表的文献中但尚未被人们认识或发觉的知识片段间的逻辑联系,提出知识假设,以便专业研究人员进一步证实。本文作者自行编写了切词、词频统计、权重计算等相关程序,分别以文献篇名和摘要为对象,采用 Swanson 的闭合式方法在中文文献中挖掘,验证了非相关文献知识发现法。将非相关文献知识发现方法引入中文科技文献领域,证明这一方法在中文文献中同样适用。希望这一方法能引起国内研究者的关注,设计出适合中文数据库的知识发现系统。

1　非相关文献知识发现方法概述

1986 年,芝加哥大学 Don R. Swanson 教授发现雷诺氏病和食用鱼油之间存在着隐含的逻辑关联。1988 年,他对偏头痛和镁缺乏进行研究,提出了第二个例证。后来,他又进行了深入细致研究,如静氨酸和生长调节素的血液水平的研究,消炎痛和 Alzheimer 疾病的研究,雌激素和 Alzheimer 疾病的研究,磷脂酶和睡眠的研究等,更进一步论证了他的非相关文献的知识发现思想[1-2]。

Swanson 的基于非相关文献的知识发现的主要思想为:分别用 A、B、C 表示 3 个不同的抽象概念,在已发表的一些文献中表明 A 与 B 存在联系,而另一些文献表明 B 与 C 存在联系;如果把两类文献放在一起研究,通过逻辑递推关系,可推断 A 与 C 之间可能存在联系,但在所有已知文献中未发现对 A 与 C 之间关系的揭示,即这种联系此前在任何文献中都没有记载。Swanson 试图通过对已有文献的挖掘,发现隐匿于文献中的未被揭示的 A 与 C 之间的联系,并根据这种联系建立一定的知识假设来指导科学实践。Swanson 教授将这种模式称为 ABC 模式,可表示为:如果 A→B 和 B→C 存在,则 A→C 可能存在[3]。

他在以后的研究中一直致力于探寻 A 与 C 之间的联系。他指出知识发现过程可以分为两个步骤:形成假设和检验假设[4]。2001 年,Weeber 正式将形成假设的过程称为"开放式方法",而将检验假设的过程称为"闭合式方法"[5]。

所谓开放式方法,主要用于科学假设的形成阶段,即 A→C 的过程。以感兴趣的主题 A 为初始点,发现中间集合 B,通过中间集合 B 与文献集合 C 之间的关系,确定 A 与 C 之间的关联[6]。开放式的知识发现过程可能是为疾病寻找一种新的治疗方法,或为药物寻找新的靶标。

所谓闭合式方法,主要用于科学假设验证阶段。如果研究人员已经通过开放式知识发现方法形成了假设,他就可以通过分析处理大量文献来详细论证自己的假设。具体做法是从 A、C 两端同时开始检索,努力寻找共同的中间词,产生相互交叉的词汇集合 B。

1991年,为了更好地实践ABC的知识发现模式,Swanson教授与他的合作者Neil R. Smalheiser设计了一个应用于生物医学文献知识发现的人机交互计算机软件系统——Arrowsmith,用于分析研究非相关的互补文献。该系统通过采用禁用词表过滤、语义网络过滤、频率过滤、日期过滤、排序等数据处理方法找出A、C所有的隐藏关联B,引导知识发现[7]。

2 在中文文献中验证Swanson知识发现方法的意义

Swanson的非相关文献知识发现法对学科体系中应用理论和技术有指导作用。他的方法得到科学假设,有很强的方向性和目的性,研究人员会节省更多的时间和精力,他们的研究工作会更加有效和富于创造性[8]。该方法在揭示文献间隐含关联方面所取得的成功向人们展示了一种新的情报研究方法,提供了一条新的研究途径[9]。非相关文献知识发现法在科学研究,尤其是医学研究上会有愈来愈广阔的应用前景。

近20年来,研究人员对非相关文献知识发现方法的研究都是建立在对外文文献进行挖掘的基础上,而且文献来源大多是MEDLINE医学文献数据库。国内对该方法的研究还处于起步阶段,2000年才开始出现对Swanson方法及Arrowsmith软件的介绍,目前已有的文献多数是对Swanson方法的探讨和评价,实践方面只有网络版Arrowsmith软件的利用,且都是基于外文文献,很少有关于中文文献知识发现过程方面的研究。由于英语与汉语在语法结构和语义理解上都存在一定的差别,Swanson方法目前还不能直接应用于中文科研领域,进一步探讨Swanson的知识发现方法在中文文献中的实用性就显得十分必要。此次验证的目的就是通过在中文文献中检验Swanson关于"雷诺氏病和鱼油"、"偏头痛和镁缺乏"之间存在联系的假设,来确认Swanson的非相关文献知识发现方法是否能够应用于中文文献,并且进一步扩展Swanson的发现,将这一方法引入中文文献领域。

3 在中文文献中验证Swanson知识发现方法的基本思路和文献源选择

本文采用闭合式方法验证Swanson的非相关文献知识发现法。若A和C两类文献从没有被共同引用,并且也不相互引用,则称这两类文献是相互独立的,即非相关文献。假设按照Swanson的知识发现方法(开放式方法)已经得到A和C之间有一定隐性逻辑关联的假设,或者,专业人员根据经验猜测A、C之间可能存在逻辑关系,但现有文献并未直接揭示这种关联。现在要做的就是在中文文献中分别以A和C为出发点,通过一系列数据和内容分析,努力寻找共同的中间词B,验证这一假设在中文文献中是否成立。如果能找到共同的中间词B,则说明A→B与C→B同时成立,如此证明A→C成立。如果对于"雷诺氏病、鱼油"、"偏头痛、镁缺乏"两对知识概念均能按上述方法找到共同的中间词,则证明Swanson的非相关文献知识发现方法适用于中文文献。

为了保证文献的权威性和学术性,本文以中国期刊全文数据库作为实验数据库。中国期刊全文数据库是目前最大的大型集成化动态中文全文信息资源。本次验证选取的文献是1979年至2006年中国期刊全文数据库医药卫生大类的期刊文献。由于运算量太大,到目前为止,Swanson的非相关文献知识发现的处理对象仅限于文章篇名。本文用闭合式方法仅验证两对知识概念,运算量大大减小,因此可以用文摘作为处理对象。从中国期刊网中检索并下载A和C两类文献的篇名与摘要,作为验证Swanson闭合式非相关文献知识发现法的测试集。

4 验证方法的设计

整个验证过程共分为7个步骤。

步骤1:检索文献数据库,分别得到与A、C相关的文献集合。

分别以A、C为检索词,从中国期刊网全文数据库中按主题途径检索出与词A、C相关的文献,得到相应的文献集合{A}和{C}。由于要验证的两对知识概念间的关系已被Swanson所揭示,可能已经存在专门讨论它们之间关联关系的文献,这类文献必须作为误检从{A}、{C}中删去。将去掉误检后的两类文献的篇名和摘要分别保存下来存入数据库,并统计{A}、{C}的文献数量。

步骤2:词切分。

汉语句子中的词语之间既无空格,又无特殊的间隔标志,对文献进行处理时一个不可回避的问题就是切分词。对篇名和摘要进行切分词时采用主题词表法。设计一个用于切分词的程序,然后将医学主题词表、停用词表和半停用词表导入程序中进行切分,并将切分后的词(关键词)保存到数据库,去掉停用词和一些无意义的词,得到一个初步的B词列表。

步骤3:合并同义词。

根据同义词表,通过程序对关键词中的同义词进行合并。本文依据的同义词表是哈尔滨工业大学信息检索实验室的HIT-IRLab同义词词林和医学主题词表。

步骤4:统计词频。

设计一个统计词频的程序,对关键词进行统计。不仅要统计每个关键词在所有记录中的词频,还要统计包含这个关键词的记录数。

步骤5:计算权重。

研究者提出了好几种非相关文献知识发现的权重计算公式,其中就有Swanson教授本人的公式。设{AB}为文献集合{A}中包含有某个B词的文献组成的子集,{BC}为文献集合{C}中包含有某个B词的文献组成的子集,Swanson以篇名为对象的闭合式方法B词权重计算公式为[10]:

$$weight = 100 \times ncoml/(nAB \times nBC)$$

此公式中,nAB为{AB}的记录数,nBC为{BC}的记录数,ncoml为{AB}和{BC}相同的篇名关键词数。

仿照Swanson的公式,我们可以给出以摘要为对象的闭合式方法B词权重计算公式:

$$weight = 100 \times ncom2/(nAB \times nBC)$$

该式中,ncom2为{AB}和{BC}相同的摘要关键词数。

按照上述公式计算B词的权重,并根据所得权重的大小对B词进行排序。

步骤6:确定阈值,选取合适的中间词。

中间词B集合中的词汇需要一定的筛选和过滤机制,权重阈值的设定非常关键。权重过高的词表明A、C在B领域的研究已经比较成熟,进行新的知识发现的可能性不大;而如果B的权重过低,则可能包含偶然性的因素较大,进行研究的价值不大。具体的范围可根据B词列表的结果选定,目前还没有形成定论。很多研究人员根据自己研究需要和词汇检索结果情况设定不同的阈值。确定阈值后,对阈值范围内的词还要进一步筛选和排除,去掉常用的动词和形容词,这些词没有实际的研究价值,无实际研究意义的名词也没有必要保留,研究者可以根据自己具体的研究需要进行筛选,确定中间词集合B。

步骤7:分析结果,得出结论。

整理并分析得到的中间词集合B,然后返回到包含B词的集合{A}、{C}中的对应文献中,阅读文献的具体内容,分析其中是否有逻辑关联,得出验证的结论。

5 Swanson关于雷诺氏病和鱼油的假设的验证

在Swanson关于雷诺氏病和食用鱼油的假设中,A代表食用鱼油;B代表血液和循环系统的一系列变化,即血液粘稠的降低,血小板聚集的降低,还有血管收缩的减少;C代表雷诺氏病,未知的外部循环混乱和相对抵抗力的改善。由已知文献可得到两个结论:(1)食用鱼油可以引起特定的血管变化,即A引起B(A→B);(2)同样的血管变化可以改善雷诺氏病,即B引起C(B→C)。由ABC模式得出假设:A引起C,即雷诺氏病与鱼油之间有一定联系,食用鱼油可能对雷诺氏病有治疗作用[11]。

以主题作为检索入口,检索出关于雷诺氏病和鱼油的文献,去掉其中误检的文献,得到与雷诺氏病有关的文献有493篇(集合A),与鱼油有关的文献有617篇(集合C)。对集合A和集合C的文摘分别进行切词、排除停用词、合并同义词、统计词频、计算权重等操作,得到包含389个词的候选B词集合。

设定关键词集合B的阈值为200~350,这样就大大缩小了B词的数量。在阈值范围内的B词有一些是无实际研究意义的词,如复合、结合、外、作用、型、观察、影响、分析、分类、分离、合并、测定、实验、症状、体、系统、化学、病、不良反应、机制、剂量等,可以去掉。经过层层筛选后,集合B中的中间词共有51个(见表1)。

表1　雷诺氏病—鱼油中间词集合B

Bword	Weight	Bword	Weight
葡萄糖	350	流行病学	230
肝硬变	333	**动脉粥样硬化**	226
前列腺素	329	**血液循环**	225
血管收缩	325	病理学	225
免疫抑制剂	325	心律失常	225
血液粘度	322	支气管哮喘	225
前列环素	317	**静脉**	224
增殖	317	解毒	217
内皮素	317	微循环	217
甘油	300	**血液**	209
颌下腺	300	**血清**	201
血浆	280	**凝血**	200
稠度	280	肝素	200
血液流变学	270	肺癌	200

续表

Bword	Weight	Bword	Weight
合酶	267	拮抗	200
缺血	261	氨气	200
肌纤维	250	链激酶	200
受体拮抗剂	250	放射病	200
羟色胺	250	基底膜	200
毒理学	250	淋巴细胞	200
结肠	250	肌细胞	200
中风	247	细胞毒	200
硫化物	244	带状疱疹	200
血红蛋白	240	妊娠	200
浓度	238	抗疟药	200
血压	238		

从实验数据可以看出，血管收缩、血液粘度、血浆、血液流变学、血红蛋白、血压、动脉粥样硬化、血液循环、静脉、血液、血清、凝血（表1中黑体表示的词）等12个包含关于血液和血液循环系统变化的中间词进入我们的视野，可选作有效B词。按照这些中间词返回到文献中，通过阅读文献、分析文献内容可以发现，鱼油有利于心血管功能，可以降低血液粘度，促进血液循环，提高血浆蛋白的含量，可以用于心脑血管疾病的防治，而血液粘度的降低、血小板聚集程度的降低以及血管收缩的减少可以改善雷诺氏病。由此可见，鱼油和雷诺氏病之间确实存在隐性的逻辑关联，两类相互独立的文献通过血液系统的变化联系到一起，食用鱼油可能对雷诺氏病有治疗作用。因而证明了Swanson关于雷诺氏病和鱼油的假设在中文文献中也同样适用。

6 Swanson关于偏头痛与和镁缺乏假设的验证

镁缺乏和偏头痛之间的关联是Swanson更为复杂的一个发现。在关于镁缺乏和偏头痛的假设中，A代表镁缺乏，C代表偏头痛，B代表两者之间的关联词。Swanson发现了如下11种镁影响偏头痛的方式：①脑皮层压抑的扩散与偏头痛有关联，镁可以减少动物的脑皮层压抑的扩散；②癫痫症和偏头痛有关联，镁缺乏可能增加癫痫症的发病率；③P物质（神经肽的一种）可能引起偏头痛，镁对P物质的活动有反作用；④偏头痛患者有很高的血小板聚集，镁可以抑制血小板聚集；⑤偏头痛患者的血小板对血液中的复合胺释放非常敏感，镁可以抑制血液中的复合胺引起的血管收缩；⑥钙拮抗剂已经成功地用于治疗偏头痛，镁是钙的天然拮抗剂；⑦压力和人的A型行为与偏头痛有关，压力和A型行为导致身体缺少镁；⑧过度的血管收缩可能增加偏头痛的易感性，镁可以减少动脉血管收缩；⑨反常的低环前列腺素的释放可以加重血管收缩对偏头痛的影响，镁可以增加环前列腺素的形成；⑩偏头痛可能引起大脑血管的炎症，镁有抑制炎症的特性；⑪大脑缺氧可能是偏头痛的关键因素，镁可以保护大脑避免缺氧的损害[12]。

以主题词作为检索入口，从中国期刊网中检索出关于镁缺乏和偏头痛的文献，去掉其中误检的文献，得到与镁缺乏有关的文献388篇（集合A），与偏头痛有关的文献5768篇（集合C）。分别下载集合A与集合C的文摘，同时经过分词、去掉停用词、合并同义词、词频统计、计算权重等步骤，得到包含642个词的候选B词集合。

设定B词集合的阈值为125～400，再去掉虽在阈值范围内但没有实际研究价值的词，得到共有92个中间词的集合B（见表2）。

表2 镁缺乏—偏头痛中间词集合B

Bword	Weight	Bword	Weight
脑缺血	396	酵解	200
脑水肿	375	血脉	200
醛固酮	367	低镁血症	197
脑血栓形成	363	静脉	195
高脂血症	363	**心血管疾病**	193
肌炎	350	苛性碱	190
神经激素	350	叶酸	188
脂类	350	血压	186
血药浓度	350	**儿茶酚胺**	183
钙通道	337	苯	182
血栓形成	331	低血糖	180
半胱氨酸	325	横膈膜	180
白细胞介素	300	松果体素	178
高碳酸血症	300	癌症	178
脑脓肿	300	中枢神经系统	177
去甲肾上腺素	300	**癫痫**	176
降钙素	300	**钙化**	175
动脉粥样硬化	290	氟化	175
低分子右旋糖酐	260	甲状腺机能亢进	175
硫酸镁	257	氯化	173
再生障碍性贫血	250	羟	173
十二指肠溃疡	250	**大脑皮层**	171
癫痫发作	247	葛根	171
缺氧	247	颌下腺	169
大脑	244	白粉病	168
血脑屏障	242	肾上腺	167

续表

Bword	Weight	Bword	Weight
凝血	238	溶血	167
缺血性脑血管病	231	肺静脉	167
炎症	226	磷酸	166
支气管哮喘	212	贫血	165
脑脊液	209	**血管系统**	163
葡萄糖	209	视网膜	161
血液	209	凝血时间	160
肝硬变	207	氟	151
胆固醇	204	肿瘤	151
激素	204	萎缩性胃炎	150
性激素	204	血液循环	150
哮喘	201	雌醇	150
血清	200	咪唑	150
脑磷脂	200	山莨菪碱	144
凝血酶原时间	200	咖啡因	140
肾病综合征	200	**胺**	135
丙酮	200	黄疸	130
阿魏酸	200	**抗癫痫药**	129
多巴	200	过敏反应	125
菌血症	200	糖皮质激素	125

表2中的钙通道、降钙素、癫痫发作、缺氧、缺血性脑血管病、炎症、心血管疾病、儿茶酚胺、癫痫、钙化、大脑皮层、血管系统、胺、抗癫痫药等中间词都反映了假设中镁缺乏和偏头痛的隐含关联。按照这些中间词返回到文献中,通过阅读文献、分析文献内容可以发现:镁缺乏可能增加癫痫症的发病率,而癫痫症与偏头痛之间也已经证明有一定联系;镁可以抑制血小板聚集和血液中的复合胺引起的血管收缩,这些都与偏头痛有关;钙拮抗剂可以用于预防偏头痛,而镁是钙的天然拮抗剂;镁可以保护大脑避免缺氧,而大脑缺氧可能是偏头痛的关键因素等等。可见,镁缺乏和偏头痛之间确实存在隐性的逻辑关联,证明了 Swanson 关于镁缺乏和偏头痛的假设在中文文献中是成立的。

7　结论

本文通过在中文文献中验证 Swanson 的知识发现方法,证实这种方法在中文文献中也同样适用,他提出的非相关文献知识发现规律也可以应用于中文科学领域。

到目前为止,我国对 Swanson 知识发现方法的研究还处于起步阶段,未见有关开放的知识发现方法过程方面的研究。但是这一方法已经开始引起国内研究人员的关注。希望在不久的将来,国内研究者能将 Arrowsmith 应用于中文医学,甚至中医数据库,设计出适合中文数据库的 Arrowsmith 系统,创造出有中国特色的非相关文献知识发现法。

参考文献

1　Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med,1986,30(1)
2　Swanson. Migraine and magnesium:Eleven neglected connections. Perspect Biol Med, 1988,31(4)
3　Swanson. Smalheiser. An interactive system for finding complementary literature:a stimulus to scientific discovery. Artificial Intelligence,1997,91:183~203
4　郝丽云,郭启煜. 非相关文献知识发现研究进展. 情报学报,2006(3)
5　Weeber,Klein,De Jong-van den Berg. Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries. J Am Soc Inf Sci Technol,2001,52(7)
6　安新颖,冷伏海. 基于非相关文献的知识发现原理研究. 情报学报,2006(1)
7　董风华,兰小筠. 基于文献的知识发现工具 - Arrowsmith. 情报杂志,2004(5)
8　许建阳,马明,王发强. Swanson 的非相关文献知识发现法对医学发展的思考. 医学与哲学,2003(8)
9　荣毅虹,梁战平. 基于文献的发现. 情报学报,2002(4)
10　Swanson, Smalheiser. Torvik. Ranking Indirect Connections in Literature-based discovery: The Role of Medical Subject Headings (MeSH). Journal of the American Society for Information Science and Technology, 2006,57(11)
11　Swanson. Two Medical Literatures that are Logically but not Bibliographically Connected. Journal of the American Society for Information Science,1987,38(4)
12　Swanson. A Second Example of Mutually Isolated Medical Literatures Related by Implicit, Unnoticed Connections. Journal of the American Society for Information Science, 1989, 40(6)

黄水清　南京农业大学信息科学技术学院教授。通讯地址:江苏南京中山门外卫岗南京农业大学信息科技学院。邮编 210095。

熊　健　南京农业大学信息科技学院2006级硕士研究生。通讯地址同上。

李志燕　南京农业大学信息科技学院2004级硕士研究生。通讯地址同上。

(来稿时间:2006-12-28)