# 基于生物医学文献的知识发现研究

黄凯峰,何洁月
(东南大学 计算机科学与工程学院,江苏 南京 210096)

**摘 要**:随着生物医药文献的快速积累,利用文本挖掘技术处理海量的科技文献,从而发现生命科学领域新的知识,已成为当前数据挖掘和人工智能领域研究的热点。从Swanson最早提出基于生物医学文献的知识发现方法到现在,许多研究人员投入到这个新兴的领域中。对基于生物医学文献的知识发现的研究内容、研究方法以及成果进行了系统的分析和阐述,对不同的研究方法在文本挖掘过程中的优劣进行了比较,对基于生物医学文献的知识发现的发展趋势进行了展望。

**关键词**:生物信息学;基于文献的知识发现;文本挖掘;知识发现
**中图分类号**:TP182;Q811.4 **文献标识码**:A **文章编号**:1673-629X(2008)02-0062-04

# Survey of Biomedicine Literature-Based Discovery

HUANG Kai-feng, HE Jie-yue
(School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

**Abstract**: With the rapid accumulation of biomedical literature, using text mining methods to process huge literature for discovery of novel scientific knowledge has been the foci in data mining and artificial intelligence research fields. After Swanson put forward the literautre-based discovery, many researchers flung themselves into this new field. This paper introduced these methods in detail, including main issues, accomplishment and compared the key methods from text mining perspectively. Finally, several challenging researching problems are identified.

**Key words**: bioinformatics; literature-based discovery; text mining; knowledge discovery

## 0 引 言

信息技术的飞速发展,积累了海量的数据,例如,当今世界上最大也是最权威的生物医学文献数据库MEDLINE就拥有超过1200万篇的文献,并且仍以每年40万条记录的速度在增加。这么多的数据,已经超出了个人所能处理的限度。同时,如此多的文献资源,也为科研人员运用数据挖掘和文本挖掘技术,发现隐含的、有价值的知识提供了机会。

文本挖掘[1](Text Mining)是指从大量文本数据中提取出可理解的、未知的、最终可用的知识的过程。它是一个交叉的研究领域,涉及了数据挖掘、信息检索、自然语言处理等多个领域的内容。因为大多数的生物信息都被保存在文本中,最近在NATURE[2]上发表的文章就指出,文本挖掘由于其对于文本的知识发现能力,可以在概念生物学(Concept Biology)中发挥支撑作用。同时,一些科研人员利用文献挖掘工具,发现了许多对人类有用的知识,例如:鱼肝油可治疗雷诺式症[3]、蛋白质之间的相互作用力[4]等。随着生命科学文献和注释性数据库的快速积累,相信通过文本挖掘技术,更多的医学知识将会被发现。

笔者系统地综述了生物知识发现的一个重要研究领域,即基于文献的知识发现的定义和研究现状,对各主要方法的文本挖掘过程进行了比较。

## 1 基于生物文献的知识发现定义

1986年,华盛顿大学的Swanson教授提出了基于文献的知识发现(Literature-based Discovery)的理论,指出非相关的生物文献中可能隐含着大量的不为人知的科学知识(Undiscovered Public Knowledge, UPK)[3]。

Swanson基于文献的知识发现定义如下:如果有两类文献集AL和CL,其中AL主要讨论了概念A和概念集B之间的关系,而CL则讨论了概念C和概念集B之间的关系,但是却没有任何的文献直接讨论过A和C的关系,那么A与C之间通过共同的桥梁B,隐含地存在某种关系,这就可能是一个新的科学发现。这时的AL和CL就称为非相关(Non-interactive)且

**收稿日期**:2007-05-15
**作者简介**:黄凯峰(1983-),男,江苏无锡人,硕士研究生,主要研究方向为数据挖掘、文本挖掘、生物信息学;何洁月,教授,研究方向为数据库技术、数据挖掘、生物信息学、信息集成等。

互补(Complementary)的文献,而概念集 B 则称为中间概念(Intermediate Concept)。

具体的过程可见图 1。

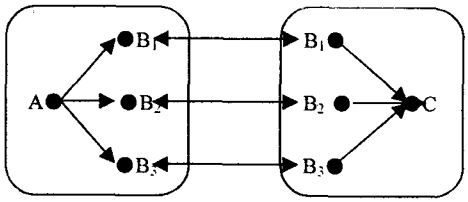


图 1　基于文献的知识发现

## 2　基于文献的知识发现的主要方法

Swanson 等人在基于文献的知识发现领域的出色工作,吸引了许多其他的研究者的注意,这其中包括了医学、生物学、情报学和计算机学的研究者。同时,各种不同的方法被应用到这个领域,例如统计学、信息检索、数据挖掘和应用本体等方法。笔者基于文献的知识发现按不同的研究方法进行了分类和介绍。

### 2.1　基于统计学的方法

Swanson 教授[3,5,6]是基于文献的知识发现这一理论的创始人,他使用的方法主要是基于单词的词频统计。以鱼肝油为例,他首先从 MEDLINE 数据库中,下载标题中含有鱼肝油的文献集 AL,统计出和鱼肝油共现的单词的频率,经过去除停词表、通用词汇和排序后,形成词表 B。然后再对 B 中的每个词以相同的方法搜索并下载文献集 CL 后进行分析,得到与 A 具有潜在联系的词表 C。基于此原理,Swanson 发现了许多对人类有益的知识。例如,鱼肝油对于雷诺氏症的治疗作用[3],镁的缺失会引起偏头痛,某些病毒可以成为潜在的生化武器等等,这些发现都得到了临床上的证实[7]。

Swanson 和 Smalheiser 经过多年的研究,在原始方法的基础上,适当地引入了语义概念的机制。他们所开发的 ARROWSMITH 系统[5],既能将中间概念集限定在化学物质、生物体或是地理区域等语义概念之下,又可以让用户自行地设定语义类型限制。不同的研究者,可以根据自己的研究领域,通过 ARROWSMITH 进行有目的的查询和检索,从而找出有价值的知识。

Gordon 等[8]是最早对 Swanson 的发现进行验证的研究人员。虽然基本沿用了 Swanson 的发现框架,可是他们的方法结合了医学的知识和信息检索的技术,从而可以对文本进行更深层次的挖掘。不同于 Swanson 的方法,Gordon 的分析范围在标题的基础上,加入了文献的摘要,提取的对象也从单词变成了 1 个、2 个或是 3 个词汇所组成的短语。这些变化,使得文本分析的范围得到了扩充,从而为知识发现提供了更加全面的准备。在对短语进行频率统计后,Gordon 使用了四种参数(Token frequency(tf),Document frequency(df),Relative frequency(rf),Tf * idf)来评价短语间的关联度。

### 2.2　基于关联规则挖掘的方法

关联规则挖掘[9]是数据挖掘技术中常用的方法,可以用来发现大量数据中项集之间有趣的关联。关联规则挖掘试图在数据库的事务中寻找形如 X->Y 的有趣的模式,这意味着包含 X 的事务也会同时包含 Y。兴趣度则由支持度(Support)和置信度(Confidence)来表示。

Hristovski[10]将关联规则挖掘引入了基于文献的知识发现,用来取代 Swanson 的简单词频方法。他将生物文献看作数据库中的事务,而用来代表文献内容的词则看作是规则中的项,通过设置支持度阈值和置信度阈值来产生关联的词汇。值得注意的是,不同于 Swanson 使用文献标题中的单词,Hristovski 使用了 MeSH(Medical Subject Headings)词作为分析的基本对象,MeSH 词是经过专家挑选的生物医学词汇,并被指派到文献中用来表示文献的主要内容。

和一般的关联规则挖掘所遇到的问题一样,由于有效阈值的设定很困难,所以 Hristovski 的方法会产生大量的候选规则集,并且消耗很多的计算时间,这就造成了系统的可用性下降。在关联规则挖掘中,如果阈值设置的过低,会产生大量的候选规则,而设置的过高,则有可能过滤掉许多有意义的规则。所以,合理的阈值设定以及筛选、排序机制仍然需要进一步的研究。

### 2.3　基于信息测度的方法

不同于其他人的方法,Wren[11,12]认为问题的关键,不在于寻找存在大量冗余的关联词汇,而在于如何有效地识别出词汇间具有信息的关联。两个词之间的关联度,可以通过信息测度的方法计算,不需要引入任何的领域知识。他使用互信息方法(Mutual Information Measure,MIM)来计算词的关联度。Wren 将由自然语言处理模块所提取出的生物词汇,构建其 Scale-free 图,并计算图中有直接联系的词汇之间的互信息值。其中 $MIM(A,B)=(P_{ab}/(P_a * P_b))$($P_{ab}$是 A、B 同时在文献中出现的概率,$P_a$、$P_b$ 是 A、B 分别在文献中出现的概率),最后使用一个连接的互信息方法对隐含的共享关系进行打分和排名。

Wren 尝试了平均互信息值和最小互信息值两种方法,在经过多次的试验后,他发现最小互信息法更容易识别出信息的隐含关联。

Wren 的算法从词汇间的共享联系出发,引入了信

息论中的互信息方法,通过互信息值的大小来表示关联的强度。他的方法具有领域无关的特性,可以用来推广到很多的研究领域。可是,也正因为这个原因,一些模糊的和宽泛的概念因为出现的频率很多,比较容易得到较高的排名。而且,信息论中的其他方法,如 $X^2$ 检验、数似比率(log - likelihood ratios)、z - scores 和 t - scores 等,Wren 也没有进行相应的验证。

### 2.4 基于语义的方法

以上的几种方法中,分析的单元停留在了单词或是短语上,并没有从词汇本身所表达的生物意义出发。Weeber[13,14]等人认识到,只有客观的生物概念,才能体现丰富的生物医学内涵,从而提供更好的挖掘结果。而本体(Ontology)作为一种概念化的显式说明,可以对客观存在的概念和关系进行描述,不仅是解决语义层次上万维网信息共享和交换的基础,也可以应用到文本挖掘的领域。其中,UMLS[15]是被使用较多的一个本体,它将生物概念分类在不同的语义类型中,并通过语义关系来表示语义类型之间的联系(见图 2)。

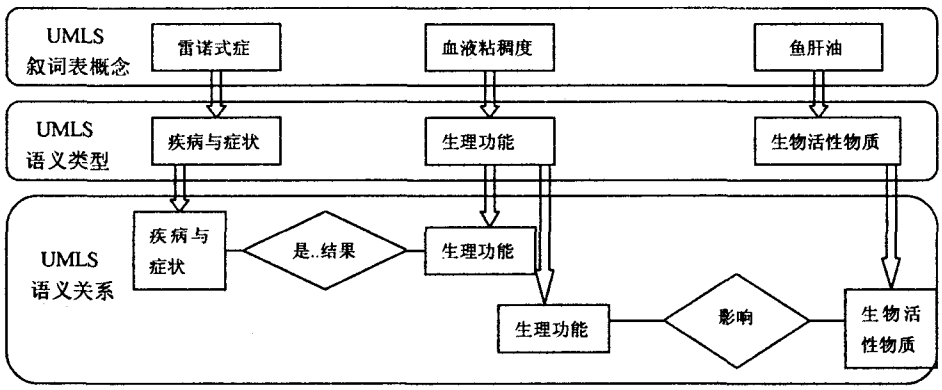


图 2 UMLS 本体结构的一个示例

在 Swanson 的理论基础上,Weeber 使用了一个两步发现模型,即首先通过开放式发现方法寻找新的科学假设,然后利用封闭式的发现方法进行验证。Weeber 设计的文本挖掘工具 DAD 系统,利用自然语言处理系统 MetaMap 将文献中的语句映射为 UMLS 中的生物概念,用概念来取代词汇作为知识发现的基础。在对概念进行统计的基础之上,DAD 系统还利用了 UMLS 中 135 个不同的语义类型和 54 种语义关系,根据概念之间客观存在的生物联系,将中间结果集进行概念簇集,并对搜索结果进行筛选。这种情况下,一些频率较少的概念通过簇集得以保存,而那些和挖掘目标毫无生物联系的概念则可以被删除。该方法超越了 Swanson 等人简单的词汇统计,实现了语义层次上的知识发现。除了成功地复制 Swanson 几个著名的发现外,Weeber 人还利用 DAD 系统找出了生姜潜在的医疗作用。

## 3 文本挖掘过程比较

与一般的文本分类和文本聚类不同,基于文献的知识发现,主要是分析生物文本中的词汇或概念之间的联系,并找出其中隐含的知识,许多研究者都对该研究方法进行了验证和改进。文中从挖掘过程入手,在"文本挖掘对象"、"文本分析单元"和"关联度量方法"等方面,对它们进行了比较(见表 1)。

表 1 基于文献的知识发现方法比较

| | Swanson | Gordon | Hristovski | Wren | Weeber |
|---|---|---|---|---|---|
| 文本挖掘对象 | 标题 | 标题和摘要 | 标题和摘要 | 标题和摘要 | 标题和摘要 |
| 文本分析单元 | 单词 | 短语 | MeSH 词 | 短语 | UMLS 概念 |
| 挖掘结果的评估方法 | 单词词频 | If, df, rf, If * idf | 支持度、适应度阈值 | 互信息阈值 | 概念频率,语义关系 |
| 自动化程度 | 需大量手工操作 | 需大量手工操作 | 需一定的手工操作 | 需一定的手工操作 | 自动化程度较高 |

* 文本挖掘对象。

初期 Swanson 的方法,由于需要大量的手工操作,文本挖掘对象限制在文献的标题。而标题有的时候不能很客观、全面地表示文本要表达的内容。Gordon、Weeber 等人将分析范围扩展到了标题和摘要,为挖掘结果提供了有效的知识保障。

* 文本分析单元。

Swanson 的文本分析单元是词,而文本中的词大多是不具有生物信息的,只是简单地进行词频统计,会产生大量的无关数据。Gordon 的方法虽然将词扩大到了短语,可是仍然没有从语义上解决问题。Weeber、Srinivasan 等人,充分利用了 UMLS、MeSH 等本体,将自由文本映射成生物概念词或 MeSH 词,从而使得分析单元可以更加客观地表示生物概念。

* 挖掘结果的评估方法。

基于文献的知识发现最重要的一点,就是如何找出隐含的、具有语义关联的生物概念,从而进行下一步的推理,这也是最为复杂的步骤。研究人员应用了大量的方法,如统计学方法、信息测度、关联规则挖掘、信息检索和本体的方法。目前统计学方法是使用最多的,但是却无法体现知识的客观性。但随着本体技术的引入,相信可以在语义的层次上进行有效的知识发现。

* 自动化程度。

最初 Swanson 的方法中,需要使用者手工地从大量的中间和最终结果集中,挑选出真正有意义的概念,

这就要求使用者拥有极强的生物和医学背景。然而,数量庞大的结果却使得人们无从下手。Weeber 等人借助 UMLS 本体的优势,在语义层次上,对挖掘结果进行自动筛选、分类,极大地减少了专家的介入。

## 4 结论与展望

随着文献资源的增加,文本挖掘技术在商业开发、科学研究领域都可以发挥巨大的作用。相对于普通的文本分类、文本聚类所提供的信息检索等功能,文中介绍的基于生物医学文献的知识发现,更加贴近人类的生活。生物医学文献中的隐含知识,可以解决一些疑难杂症,也能增长人类的健康知识。生物数据库中隐含的基因、蛋白质关系,更可以为人类基因组计划提供帮助。

目前基于文献的知识发现领域内的研究人员,在基于统计学方法的基础上,逐渐开始利用 UMLS 等本体来提高知识发现的效果。本体可以更加客观地进行知识的表示和推理,并大大减少了噪声数据和手工操作。随着本体技术的日益成熟,如何更有效地应用本体,是今后的一个重要的研究方向。另外,大多数的研究人员,一直致力于如何找出关联的概念,却忽视了如何才能更好地对于有意义的关系进行排名,特别是如何用语义对它们进行排名,让专家可以更方便地发现隐含的知识。

**参考文献:**


[1] Feldman R, Dagan I. Knowledge discovery in textual databases ( KDT ) [C]//KDD - 95. Montreal: AAAI Press, 1995: 112 - 117.

[2] Blagosklonny M V, Pardee A B. Unearthing the gems[J]. Nature, 2002, 416(6879): 373 - 374.

[3] Swanson D R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge[J]. Perspectives in Biology and Medicine, 1986, 31: 526 - 557.

[4] Daraselia N, Yuryev A, Egorov S, et al. Extracting protein interactions from MEDLINE using a full - sentence parser[J]. Bioinformatics, 2004, 20(5): 604 - 611.

[5] Swanson D R, Smalheiser N R. An interactive system for finding complementary literatures[J]. Artificail Intelligence, 1997, 91: 183 - 203.

[6] Smalheiser N R, Torvik V I, Bischoff - Grethe A, et al: Ranking indirect connections in literature - based discovery: The role of medical subject headings[J]. Journal of the American Society for Information Science and Technology, 2006, 57 (11): 1427 - 1439.

[7] DiGiacome R A, Dremer J M, Shah D M. Fish oil dietary supplementation is patients with Raynaud's phenomenon: A double - blind, controlled[J]. prospective study, American Journal of Medicine, 1989, 8: 158 - 164.

[8] Lindsay R K, Gordon M D. Literature - based discovery by lexical statistics[J]. Journal of American Society for Information Science, 1999, 50(7): 574 - 587.

[9] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//In VLDB'94. Santiago, Chile: [s. n.], 1994: 487 - 499.

[10] Hristovski D, Stare J, Peterlin B, et al. Supporting Discovery in Medicine by Association Rule Mining in Medline and UMLS [J]. Medinfo, 2001, 10(Pt 2): 1344 - 1348.

[11] Wren J D. Knowledge discovery by automated identification and ranking of implicit relationships[J]. Bioinformatics, 2004, 20(3): 389 - 398.

[12] Wren J D. Extending the mutual information measure to rank inferred literature relationships[J]. BMC Informatics, 2004, 5 (1): 145 - 158.

[13] Weeber M, Klein H, Lolkje T W, et al. Using concepts in literature - based discovery: Simulating Swanson's Raynaud - fish - oil and migraine - magnesium discoveries[J]. Journal of the American Society for Information Science and Technology, 2001, 52(7): 548 - 557.

[14] Weeber M, Vos R, Klein H, et al. Generating Hypotheses by Discovering Implicit Associations in the Literature: A Case Report of a Search for New Potential Therapeutic Uses for Thalidomide[J]. J Am Med Inform Assoc, 2003, 10(3): 252 - 259.

[15] Humphreys B L, Lindberg D A B, Schoolman H M, et al. The Unified Medical Language System: an informatics research collaboration[J]. Journal of the American Medical Informatics Association, 1998, 5(1): 1 - 11.


(上接第 61 页)

**参考文献:**


[1] 于 磊,王 浩,王 骋. RoboCup 中传球策略研究[J]. 计算机工程与应用,2004(28): 59 - 61.

[2] 郭 博,程家兴. RoboCup 仿真组的传球策略[J]. 计算机技术与发展,2006,16(2): 129 - 131.

[3] 赵 斌,李一民. RoboCup 仿真机器人足球赛研究[D/OL]. 2004 - 09 - 08. http://lsg.cnki.net/grid 20/detail.aspx? Query ID = 29&CurRec = 2.

[4] 彭 军,吴 敏,曹卫华. RoboCup 机器人足球仿真比赛的关键技术[J]. 计算机工程,2004,30(4): 49 - 51.

[5] 柳长安,刘 刚,刘春阳. 机器人足球防守算法研究[J]. 哈尔滨工业大学学报,2004(7): 952 - 953.