硕 士 学 位 论 文

基于语义和监督学习的生物医学文献知识发现

Knowledge Discovery in Biomedical Literature using Semantic Resources and Supervised Machine Learning

作 者 姓 名：周峰

学科、专业 ：计算机应用技术

学　　号 ：20809310

指 导 教 师：林鸿飞 教授

完 成 日 期：2010.11

大连理工大学

Dalian University of Technology

大连理工大学学位论文独创性声明

作者郑重声明：所呈交的学位论文，是本人在导师的指导下进行研究工作所取得的成果。尽我所知，除文中已经注明引用内容和致谢的地方外，本论文不包含其他个人或集体已经发表的研究成果，也不包含其他已申请学位或其他用途使用过的成果。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

若有不实之处，本人愿意承担相关法律责任。

学位论文题目：_____

作者签名：_____　　日期：_____年____月____日

摘　要

随着生命科学的不断发展，生物医学文献数量急剧增长。为了跟踪最新的领域研究进展，科学研究者需要阅读如此大量的文献，这使得研究工作变得非常困难。数量巨大的科学文献还会导致学科的细化，不同学科之间缺乏交流，导致不同学科之间隐含的有用知识被埋藏。Swanson最早开始基于生物医学文献的知识发现研究，通过挖掘生物医学文献中的隐含知识，形成生物医学假设来辅助生物医学工作者的工作。很多研究者投入这个领域，基于生物医学文献的知识发现已经成为文本挖掘的一个重要方向。

传统基于简单共现的方法会产生过多的目标词进而导致有效目标词的排名下降，并且在计算中会遇到选取合适阈值的问题。本文采用开放式的知识发现，提出一种新的选取连接词的方法，即引入监督学习的方法，综合选取连接词的多种特征。本文实验以Swanson 发现的老年痴呆症的连接词为特征，通过分类来选取初始词雷诺氏病和偏头痛的连接词，同时加上 UMLS 语义类型的过滤。相比于其他方法，本文选取了有效的连接词，减少了目标词的数量，最终使目标词鱼油和镁分别得到了较高的排名。另外，本文把这种方法应用在 H1N1 的知识发现研究中，进行开放式发现和闭合式发现，得到了较高的准确率和 F 值，并且对可能影响 H1N1 的物质进行了预测。

挖掘UMLS语义资源进行计算逐渐成为基于文献的知识发现的热点。通过概念的语义相似度计算事件相似度取得了比统计方法如tf*idf更好的结果。本文在概念的语义相似度的基础上，加入了概念的语义关联度，避免了事件之间语义相似度高而缺乏语义关联度，导致发现的假设不合理。本文的方法充分挖掘了UMLS中的语义资源，更加合理地计算了事件之间的相似度。通过雷诺氏病和鱼油以及偏头痛和镁的实验证明，这种计算方法取得较好的效果。

关键词：知识发现；监督学习；语义相似度；语义关联度

Knowledge Discovery in Biomedical Literature using Semantic Resources and Supervised Machine Learning

Abstract

Nowadays, the amount of biomedical literatures is growing at an explosive speed. Researchers struggle to maintain expertise and knowledge of developments in their fields. Dealing with the huge amount of information has led to a fragmentation of scientific literature, which promoting poor communication between specialties. Swanson initiated hidden knowledge discovery in biomedical literature and formed several hypothesis.Many other researchers have successfully replicated Swanson's discoveries, and literature based discovery has become an popular topic in text mining.

The popular methods based on co-occurrence produce too many target concepts which will lead to the decline of really relevant target concepts in ranking. This paper presents a new method for selecting linking concepts. This method uses the statistical and textual features to represent each linking concept and then classifies them as relevant or irrelevant to the starting concepts. The relevant linking concepts are used to discover target concepts. In this way, the amount of target concepts is greatly reduced and the really relevant target concepts can gain higher rankings, which helps the biomedical experts to discover potential target concepts efficiently. We also employ this method in the investigation of H1N1, which achieves better precision and F score. At last, we make a prediction of the substances which may affect H1N1.

Many researchers utilize UMLS's semantic resource in literature based discovery. Event similarity calculated by semantic similarity between concepts show better result than statistical methods such as tf*idf. But events with high semantic similarity may lead to unreasonable hypotheses due to lacking of semantic relevancy. This paper uses UMLS's semantic network to calculate semantic relevancy between concepts, and apply F score to trade-off semantic similarity and semantic relevancy. The experimental results show Fish oils and Magnesium obtains better rankings.

Key Words： Knowledge Discovery; Supervised Learning; Semantic Similarity; Semantic Revelency

目　　录

1 绪论

1.1 研究背景及现状

1.1.1 研究背景

在科学知识无限增长的当代，科学研究者往往需要通过努力地阅读来增加研究领域的专业知识。全世界有很多的科学杂志，每个杂志出版大量的文章，这使得文献数据库变得非常巨大。例如，在线数据库MEDLINE（主要面向生物医学文献），包含超过1800万篇关于生物医学文献的摘要，这些摘要来自全世界大约5400个杂志。另外，自从2005年以来，每天有2000—4000篇摘要添加进来。因此，科学研究者需要阅读如此大量的文献来了解和跟踪最新的领域研究进展。

处理如此大量的文献会导致科学文献的分支，这些分支存在于：（1） 专业：例如，生物物理学，天体物理学，数学物理学；（2） 子专业：例如，水生毒物学，蛋白质组学，分子免疫学；（3） 结构：例如，血液，细胞，脂类的研究；（4） 技术：例如，电泳疗法，质朴分析法，超显微术。Swanson认为文献的这种分支专业化会不断成为一个难题，尤其是在生物医学领域[1]。因为科学研究者会更多的与他们所在分支的其它研究者交流，而不会考虑更加宽广的范围，于是与其它领域的交流就变少了[2]。从文献的引文就可以看出，研究者更愿意引用自己领域的文献，忽略其它领域的文献。这最终会导致两类分支中的隐含的有效连接被埋藏。

传统的计算机辅助技术，例如信息检索，对于识别关联是不够的。解决的办法之一就是基于文献的知识发现（Literature Based Discovery，LBD），它是用来解决知识分支的问题，找出新的、未显示发表过的隐含连接。Swanson最早提出了从文献数据库发现新关系的观点，并在这个领域发表了多篇文章[3]。他把LBD定义为从互不相交的科学文献中寻找互补结构的过程，这个互补结构包括两个独立的部分，他们没有共同出现在同一篇文档中或是互相引用过，当结合在一起时，会产生新的、重要的推论。最终，通过知识发现过程找到的连接会帮助生物医学研究者减少工作量，并对他们的工作有一定的启发和指导作用。

1.1.2 研究现状

在基于生物医学文献的知识发现研究领域，最主要的文献数据库是MEDLINE，很多研究者在该数据库上使用了多种技术进行实验研究。他们集中在重复Swanson的发现和使用他的结果来评价自己的方法。例如Vos的发现模型关注与药物和疾病相互作用的

模式[4-5]，中间的概念可以是药品不良反应，在DAD系统中是Drug–Adverse drug reactions–Disease模式。Gordon和Lindsay采用文档数量、TFIDF等信息检索方法统计词频，部分工作中结合了生物医学专家的人工帮助来完成[6-7]。Weeber等人加入了自然语言处理工具来识别生物医学概念，并且用到了本体UMLS进行语义类型的限定，大大的减少了连接词和目标词的数量，这个过程比之前的自动化程度提高了[8]。之后，他们用这种方法研究了thalidomide这种药物潜在的用途[9]。Hristovski将关联规则挖掘引入了基于文献的知识发现[10]。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MeSH词则看作是规则中的项，通过MeSH词的共现来设置支持度阈值和置信度阈值从而来产生关联的词汇。Srinivasan提出了视图（Profile）的概念[11]，为每个MeSH连接词建立视图，视图里面的词又以语义类型归类，其本质是在语义类型过滤后进一步选取更有效的连接词，从而减少目标词的数量。Yetisgen-Yildiz和Pratt提出了使用信息检索中的准确率、召回率和MAP等方法评测知识发现研究[12-14]。Xiaohua Hu等人在传统的关联规则方法的基础上加入语义信息[15-16]，通过合理的语义关系产生候选的语义类型，得到了较好的效果。Miyanishi等人使用事件相似度从语义角度进行研究[17]，得到了比基于统计更好的结果。有些研究者使用Swanson的知识发现框架，对一些潜在的疾病的治疗和药物的作用进行了研究[9][18-22]。很多研究者把LBD应用到生物医学文献之外的领域，如Valdes-Perez使用了化学数据库[23]，Cory使用人文科学数据库找到了20世纪的一个诗人和一个古代哲学家之间的隐含关联[24]。

Swanson最初的方法非常费力费时，并且需要人工参与。后来的研究工作都在努力使得这个过程更易于执行，并且更快、更加自动化。为此，不断加入了概念抽取、结果计算以及优化输入数据的规模和类别的技术。当然，专家的角色也仍是非常重要的。这些研究趋势表明目前LBD的研究方向：首先，需要加强基于文献的知识发现的理论基础，虽然Swanson的早期工作证明了知识发现研究的有效性，但是并没有评估知识发现过程的理论基础。其次，知识发现需要一个公认合理的评测标准。最后，知识发现的自动化程度仍需继续加强。虽然完全的自动化并不是知识发现的最终目标，但是提高知识发现系统自动化的程度可以提供更快的处理速度，从而建立更大型的知识库以便进一步的研究。

1.2 本文主要工作及章节安排

本文研究内容主要包括基于监督学习和基于UMLS本体事件相似度的生物医学文献知识发现，并在这两种方法上进行了理论探讨和实践验证与分析讨论。

第一章介绍了基于生物医学文献的知识发现的研究背景，阐述了知识发现的研究现状和研究成果。

第二章说明了生物医学文献知识发现涉及到的相关资源、工具及算法，包括MEDLINE、医学主题词和医学一体化语言系统，开放式发现、闭合式发现及结合数据挖掘的方法。

第三章介绍了监督学习的方法在开放式知识发现研究中的应用，并在雷诺氏病和鱼油、偏头痛和镁的实验上验证了方法的有效性。

第四章以流行病H1N1为例，使用第三章监督学习的方法进行了开放式和闭合式知识发现，对其进行了研究和预测。

第五章使用基于 UMLS 本体的事件相似度计算方法，在语义相似度的基础上融合了语义关联度，对比了多组实验，并对实验结果进行了分析。

2 生物医学文献知识发现相关资源、工具及算法

2.1 生物医学文献及本体资源

2.1.1 生物医学文献资源

目前，基于生物医学文献的隐含知识发现使用的主要文献是美国国立医学图书馆（National Library of Medicine，NLM）提供的在线生物医学文献库 MEDLINE，这是生物医学研究领域的重要知识来源，内容涉及 1947 年至今的基础医学、临床医学、实验医学、环境和公共卫生等许多学科。这些文献的摘要来自全世界大约 5400 个杂志，包括 39 种语言[25]。MEDLINE 的一个特别之处在于所有的文档都使用 NLM 的医学主题词建立索引。MEDLINE 的结构，包括编号、题目、摘要、MeSH 等内容。图 2.1 是一个简略的 MEDLINE 结构，包含了实验用到的内容，PMID 表示一个唯一的文章编号，TI 表示文章的题目，AB 表示文章的摘要，MH 是医学主题词 MeSH。

```
PMID - 5971778
TI      - Studies on the respiratory metabolism of isolated human adipose cells.
AB      - 1. Some metabolic characteristics of fat cells isolated from 50 patients were
          examined. 2. The respiratory activity of human fat cells was of the same order
          of magnitude as cells……
MH    - Adipose Tissue/*metabolism
MH    - Carbon Isotopes
MH    - Chromatography, Thin Layer
MH    - Glucose/pharmacology
MH    - Humans
```

图 2.1 简略的 MEDLINE 结构

Fig. 2.1 A brief structure of MEDLINE article

2.1.2 医学主题词

医学主题词（Medical Subject Headings，MeSH）是 NLM 的控制语汇表，是美国国立医学图书馆编制的权威性主题词表，是用以描述主题或内容特性的 MeSH 语汇。NLM 使用 MeSH 词对 MEDLINE 文献数据库建立索引，每篇 MEDLINE 文献包含所使用的 MeSH 词都是经由专家标注的，能够反映每篇文献的中心内容[26]。通过 MeSH 词的标注，MEDLINE 文献从非结构化的数据格式转变为半结构化的数据格式。MeSH 按照字母结构和树形结构两种方式组织。最顶层的是最宽泛的概念，例如“Anatomy（解剖学）”

和“Mental Disorders（精神障碍）”，底层的是更具体的概念，处于 MeSH 的 11 层结构的下层部分，例如“Ankle（脚踝）”和“Conduct Disorder（行为失常）”。

MeSH 由主题词变更表、字顺表、树状结构表和副主题词表四部分组成，其中字顺表和树状结构表是 MeSH 的主要组成部分。

（1） 主题词变更表　主题词表是用来标引医学文献的，随着医学的发展，词表具有动态性的特点。NLM 每年都要给词表增加一些新主题词并删掉一些文献旧主题词，主题词变更表被用来反映主题词的改动情况。

（2） 字顺表　字顺表（Alphabetic List）是医学主题词表的主表。它由主题词、款目词和副主题词混合按英文字顺排列组成。

主要叙词（Major Descriptor）即主题词，用作计算机检索时的检索词，包括主要主题词、地理主题词、特征词、出版类型和类目词等。在 2010 年的 MeSH 词表里，共有 25,588 个主题词。次要叙词（Minor Descriptor）也叫次主题词，在词表中用“属”（seeunder）归入其上位主题词，而在该主题词下用“分”（XU）表示它们之间的从属关系。用次要叙词标引的文献只用于计算机检索。从 1991 年起已经停止指定和使用次要叙词。MeSH 表收入一部分不用作主题词的同义词或近义词，称为款目词（Entry Term），字顺表中用“see”参照指导读者使用正式主题词，2010 年 NLM 提供 172,000 个款目词来帮助匹配最合适的 MeSH 词，例如“Vitamin C”是“Ascorbic Acid（维生素 C）”的一个款目词。副主题词（Subheading）用于和主题词进行搭配使用检索文献，以提高主题概念的专指度，其作用实质为限定主题词的适用范围。例如，副主题词“therapy（治疗）”与疾病主题词组配，可用于综合疗法，具体如，消化性溃疡的心理疗法，用消化性溃疡/治疗。

（3） 树状结构表　树状结构表按其学科性质、词义范围的上下隶属关系，把字序表中的词分别归属在 16 个大类下，16 个大类依次用 A—N、Z 代表。它是字顺表的辅助索引，帮助确定每一个主题词在医学分类体系中的位置。一般情况下，一个词被归入一个类别并分配一个编号，但有些主题词具有一种或多种属性，则这些词同时属于两个或多个类目范畴，在其它类目亦给出相应的树状结构号，从而可以查出该词在其它类目中的位置。树状结构可以帮助研究者从学科体系中选择主题词，帮助增大或缩小检索范围，通过一个陌生主题词的上下层次的词了解该主题词，在实验中根据词的树形结构深度过滤过于宽泛的主题词等，也可以确定一个主题词的上位词和下位词的总数来确定该词的信息量。例如，MeSH 词“Migraine Disorders”的树状结构层次关系见表 2.1：

表 2.1　MeSH 词 Migraine Disorders 的树形结构

Tab. 2.1　Hierarchy structure of Migraine Disorders

MeSH 主题词	树形结构
Diseases	C
Nervous System Diseases	C10
Central Nervous System Disease	C10.228
Brain Disease	C10.228.140
Headache Disorders	C10.228.140.546
Headache Disorders, Primary	C10.228.140.546.399
Migraine Disorders	**C10.228.140.546.399.750**
Migraine with Aura	C10.228.140.546.399.750.250

2.1.3　一体化医学语言系统

UMLS（Unified Medical Language System）是美国国立医学图书馆自 1986 年起研究和开发的一体化医学语言系统，用来方便计算机系统的发展，使得这些生物医学和健康方面的专业术语变得易于使用管理和更加规范。为此，NLM 开发和发布了 UMLS 资源数据库以及相关的多种软件以供生物医学研究者更方便地使用这些资源。UMLS 资源由超级叙词表（Metathesaurus）、语义网络（Semantic Network）、情报源图谱（Information Sources Map）和专家词典（SPECIALIST Lexicon）四部分组成[27]，其中最常用的是超级叙词表和语义网络。

超级叙词表非常大，并且是多语言的词汇表，包括了生物医学和健康相关的概念，这些概念的不同形式以及他们之间的关系。超级叙词表的词汇来自各种电子版本的词库、字码集，病例、健康服务单、公共健康统计、生物医学文献索引、临床以及健康服务研究的控制语汇表。超级叙词表通过概念或含义来组织，这是为了统一具有相同含义的概念，然后识别不同概念之间的有效关系，表里的每个概念都被赋予至少一种语义网络中的语义类型。

语义网络为超级叙词表的所有概念提供一致的类别并定义了这些概念之间的一系列的有用的关系。超级叙词表定义的是概念的信息，语义网络定义了语义类型并把这些语义类型赋予每个概念，并且还定义了不同语义类型之间的关系。主要的语义类型有 Organisms（有机体），Anatomical Structures（解剖结构），Biologic Function（生物功能），Chemicals（化学物质），Events（事件），Physical Objects（物理对象）和 Concepts or Ideas（概念和观点）等。每个语义类型都有一个唯一的 ID、它的定义以及树形结构编号以表示它在层次结构中的位置。功能特性相近的语义类型可以归类为更大的语义类

型组，例如 Chemicals & Drugs，Concepts & Ideas 等。目前的发布版本中，语义网络包括 135 种语义类型和 54 中语义关系。在语义网络中，语义类型是节点，语义类型之间的关系是节点之间的连接。语义类型之间最主要的是“isa”关系，表示层次结构。另外，也有非层次结构的语义关系，共有五大类，分别是：“physically related to（物理上相关）”、“spatially related to（空间上相关）”、“temporally related to（时间上相关）”、“functionally related to（功能上相关）”和“conceptually related to（概念上相关）”。同样，每个语义关系也都有一个唯一的 ID、它的定义、树形结构编号以及可能与之有语义关系的语义类型集合。UMLS 的语义类型和语义网络的一个具体例子如图 2.2 所示：

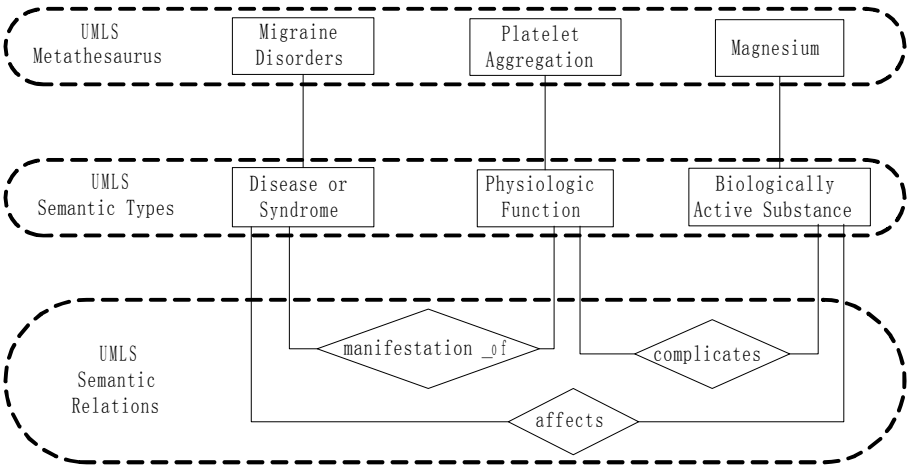

图 2.2　UMLS 示例

Fig. 2.2　An illustrative of the UMLS

UMLS 本体和 MeSH 本体是非常规范化和系统化的资源，对于生物医学研究者的工作起到了很大的帮助作用，使得众多的研究者可以对同一问题在相同的平台下交流，这无疑促进了生物医学研究的发展；同时，规范的查询系统又使得非生物医学专家可以在 UMLS 平台上进行方便的查询，并使用 UMLS 提供的多种工具，进而进行逐步深入的研究。

2.2　生物医学文献映射工具

2.2.1　MetaMap

MetaMap 是由 NLM 的 Aronson 开发的系统，该系统能够把生物医学文本映射到 UMLS 的超级叙词表里的概念，并且配置很灵活。MetaMap 主要使用基于符号的知识密集方法、自然语言处理和计算语言学方法。MetaMap 不仅应用在信息检索和数据挖掘领

域，而且是 NLM 生物医学文献建立索引的基础软件，被用来半自动和全自动地建立 NLM 的生物医学文献索引。

MetaMap 的自动文本映射过程主要包括以下步骤[28]：

（1） 将文本解析成名词短语；（2） 产生名词短语的变形词，包括名词短语中的一个或多个单词和它的变体以及它们之间有意义的组合；（3） 形成所有 Meta 入选词串集；（4） 对于每个入选的词串，计算该词串对名词短语的映射得分并排名；（5） 选择那些最高得分的，作为最佳 Meta 映射候选集。

例如，MetaMap 处理映射文本中的短语“lung cancer”时，映射会得到 8 个候选词，最终选择分值最高的“Lung Cancer”作为映射结果。

```
Processing 00000000.tx.1: lung cancer

Phrase: "lung cancer"
Meta Candidates (8):
   1000 C0242379:Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
   1000 C0684249:Lung Cancer (Carcinoma of lung) [Neoplastic Process]
   861 C0006826:Cancer (Malignant Neoplasms) [Neoplastic Process]
   861 C0024109:Lung [Body Part, Organ, or Organ Component]
   861 C0998265:Cancer (Cancer Genus) [Invertebrate]
   861 C1278908:Lung (Entire lung) [Body Part, Organ, or Organ Component]
   861 C1306459:Cancer (Primary malignant neoplasm) [Neoplastic Process]
   768 C0032285:Pneumonia [Disease or Syndrome]
Meta Mapping (1000):
   1000 C0684249:Lung Cancer (Carcinoma of lung) [Neoplastic Process]
Meta Mapping (1000):
   1000 C0242379:Lung Cancer (Malignant neoplasm of lung) [Neoplastic Process]
```

图 2.3　MetaMap 映射示例

Fig. 2.3　An Mapping process of MetaMap

2.2.2　SemRep

SemRep 是一个自然语言处理系统，通过语法分析和 UMLS 的领域知识识别出自由文本中的实体，用来提取生物医学文献中的语义假设。该系统主要使用了专家词典和词性标注以及 MetaMap 的映射结果。

对于给定的句子：Mycoplasma pneumonia is an infection of the lung caused by Mycoplasma pneumoniae.

首先，根据词和短语的不同成分把句子分成块: