数字图书馆

# UMLS 及其在智能检索中的应用*

白海燕 王 莉 梁 冰
(中国科学技术信息研究所 北京 100038)

【摘要】调研 UMLS 构成和建设特点,重点研究 UMLS 在检索方面的应用实例,分析归纳 UMLS 在语义化、智能化检索方面的功能设计、实现方法与实际效果,以期为基于集成式知识组织系统的智能检索应用的场景功能设计、技术开发和实现,提供借鉴和参考。UMLS 在智能检索中的应用主要包括:(1)扩展检索,主要有同义词扩展、等级结构扩展和词组切分扩展等方法;(2)语义检索,基于概念和概念之间的关系进行检索和结果内容表达;(3)问答式检索,包括问题分析、文献检索、语句提取、答案生成和语义聚类。

【关键词】UMLS 智能检索 扩展查询 语义检索 问答式检索
【分类号】TP391

## UMLS and Its Application in Field of Intelligent Retrieval

Bai Haiyan Wang Li Liang Bing
(Institute of Scientific & Technical Information of China, Beijing 100038, China)

【Abstract】This paper mainly investigates UMLS's composition and characteristics, emphasizes on application cases of UMLS in field of intelligent and semantic retrieval, and analyses the scene designs, technology realization and application effects of these cases, in order to provide references for design and development of intelligent retrieval system based on integrated knowledge organization systems. The application of UMLS in field of intelligence and semantic retrieval mainly includes: (Ⅰ)query expansion, such as synonym expansion, hierarchical structure expansion and phrase segmentation expansion; (Ⅱ) semantic retrieval, based on concepts and relationships between concepts; (Ⅲ) question - answering search, including question analysis, document retrieval, sentence selection and semantic clustering.

【Keywords】UMLS Intelligent retrieval Expand query Semantic search Question answering retrieval

## 1 引 言

统一医学语言系统(Unified Medical Language System,UMLS)是美国国立医学图书馆(National Library of Medicine,NLM)于 1986 年开始建设的一体化医学知识语言,具有集成性、跨领域和工具化的特点。UMLS 在信息检索(Information Retrieval)、自然语言处理(Natural Language Processing)、电子病历(Electronic Patient Records)、健康数据标准(Health Data Standards)等方面[1]得到了广泛的研究和应用。NLM 应用 UMLS 的系统和项目主要有 PubMed®,提供对 Medline 和其他相关数据库的免费检索;NLM Gateway,提供对 NLM 多个系统的集成检索,包括 Medline、OLD Medline、LocatorPlus、PubMed、AIDS Meetings、HSRProj 和 MedlinePlus 等;ClinicalTrials. gov 使用 UMLS

收稿日期:2012-03-09
收修改稿日期:2012-04-13
*本文系国家"十二五"科技支撑计划项目"面向外文科技文献信息的知识组织体系建设与应用示范"(项目编号:2011BAH10B05)的研究成果之一。

自动增加检索词的相关词;Indexing Initiative 利用 UMLS 开发自动标引系统。其他机构开发和研究包括应用于 NLM 的项目有美国国家癌症研究所的 Enterprise Vocabulary Services 和美国卫生保健研究与质量管理局的 National Guidelines Clearinghouse 与 National Quality Measures Clearinghouse 等[2]。据 UMLS 的年度统计[3],UMLS 最主要的应用方向包括:术语研究(约占所有应用的 53%)、术语映射(约占所有应用的 35%)和创建本地术语(约占所用应用的 33%)等;其他应用包括:信息标引和检索(31%)、自然语言处理(21%)等。

本文主要调研 UMLS 构成和建设特点,重点研究 UMLS 在检索方面的应用实例,分析归纳 UMLS 在语义化、智能化检索方面的功能设计、实现方法与实际效果,以期为基于集成式知识组织系统的智能检索应用的场景功能设计、技术开发和实现,提供借鉴和参考。

## 2 UMLS 的构成与特点

UMLS 的目标是力图使计算机系统能够理解生物医学和健康语言。因此,NLM 发布了 UMLS 知识源(数据库)和相关软件工具(程序),供医学信息学领域的信息系统开发人员和信息研究人员使用[4]。UMLS 包括以下 4 大部分:

(1)超级叙词表(Metathesaurus),是 UMLS 知识源的核心,由来自各种受控词表的概念和术语以及它们之间的关系所构成;

(2)语义网络(Semantic Network),是对超级叙词表概念的分类和分类之间的关系;

(3)专家辞典(SPECIALIST Lexicon),是一个词典信息库,用于自然语言处理;

(4)支持性的软件工具,各种利用 UMLS 的工具和程序[5]。

也有人将 UMLS 分为三大部分,即将上述第(3)、(4)两部分结合在一起[3]。据统计,超级叙词表是最经常被使用到的知识源,约占总使用量的 94%;其次是语义网络和专家词典与工具,约占使用量的 28%;三项都使用的用户占全部用户的 19%[3]。无论划分为 4 个部分还是 3 个部分,这几个部分既可以同时组合使用,也可以各自独立使用。

### 2.1 超级叙词表

超级叙词表(Metathesaurus)是 UMLS 的基础与核心,具有以下特点[1,6]:

(1)来源的广泛性、异构性与多语言性

超级叙词表是 UMLS 构成的基础。截至 2011 年的最新版本(2011AB),超级叙词表包含有 260 万个概念和 860 万个唯一概念名称,这些概念来源于 161 个词表源,其类型包括主题词表、分类系统、标题表、代码表、本体等,涉及 19 个语种。从当前受控词表集成的规模来看,UMLS 具有空前的广泛性、异构性和多语言性[7]。

(2)建设的开放性和可持续性

UMLS 超级词表的概念体系是一个不断累积建设的过程,1993 年,它的来源词表只有 15 个,2007 年增至 136 个来源词表,17 个语种[8]。UMLS 具有良好的维护和更新机制,包括词表新增、词表版本更新、错误修正等。NLM 网站的 What's New, Updated Sources 和 Release Documentation 的统计部分发布 UMLS 的更新情况。

(3)以概念为核心的字串 - 术语 - 概念的组织方法

概念是超级叙词表的组织核心。同一概念在不同的词表中有不同的表达,即使同一表达,也可能有不同的词形。UMLS 采用术语组织(Terms Organize)的方法,将表达同一事物的不同来源的不同表达集中在一起,形成一个概念,同时选择一个较为通用的词作为优选词(Preferred Term)来表达这个概念,并对这个概念分配一个概念唯一标识(Concept Unique Identifier),这一标识是不变的,本身也是无意义的代码。来自各个词表的同义词即术语,也会被分配一个唯一标识 LUI (Lexical Unique Identifier)[9,10]。因此,超级叙词表的概念组织模型为字串 - 术语 - 概念,如图 1 所示:

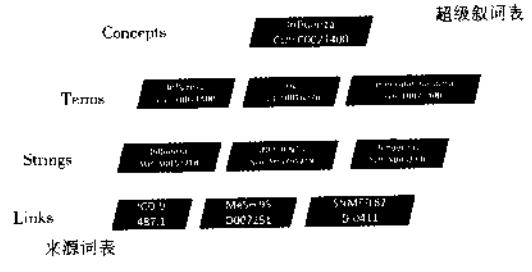


图 1 超级叙词表的概念组织模型示意图[11]

### 2.2 语义网络

语义网络由两部分组成:语义类型(Semantic Types)和语义关系(Semantic Relationships)。语义类型

是概念的范畴分类，超级叙词表中每一个概念至少要被分配一个语义类型，语义关系则是语义类型之间的关系[12]。

目前的语义类型有135个，可分为实体(Entity)和事件(Event)两大类[13]。实体指物理对象，如生物、解剖结构、物质、制品等；事件是社会活动，如行为、活动、研究过程等。语义类型是分层次的，因此具有等级关系即is-a，除此之外，语义类型之间还存在各种相关关系，如：物理上相关(Physically-related-to)，空间上相关(Spatially-related-to)，功能上相关(Functionally-related-to)，时间上相关(Temporally-related-to)和概念上相关(Conceptually-related-to)。UMLS定义了包括is-a在内的共54种语义关系，语义类型可以看成是有层次结构的"节点"，而语义关系将这些节点连成网络。

### 2.3 专家辞典及工具

专家辞典(SPECIALIST Lexicon)收录常见的英语单词、生物医学术语和出现在Medline、UMLS Metathesaurus中的术语。每个词条记录均详细描述自然语言处理系统所需要的词典信息，包括句法、形式和结构的拼写信息，同时提供词典工具和程序供超级叙词表和专家词典确定英语词汇的范围以及识别生物医学术语和文本中词的词形变异，是进行检索、标引和词汇处理的有力工具[14]。词条目可以是单个单词或多个单词组成的术语，相应的记录包括4个组成部分：基本形式、词类、唯一性标识符以及任何现成可用的拼写形式。

专家辞典提供的自然语言处理工具如表1所示[15,16]：

表1 专家辞典工具

| 工具名称 | 主要用途 |
|---|---|
| Lexical Access Tool | 提供对专家辞典的获取，输出格式为文本和XML格式 |
| SPECIALIST Lexical Tools | 使用专家辞典中的信息和其他数据，生成适用于标引和自然语言处理的词或术语的词法变异 |
| SPECIALIST Text Tools | 帮助用户将自由文本分解为词、术语、短语、语句和段落 |
| Text Categorization | 提供范畴划分和较高层次聚类的工具，可用来划分文本、标引内容、检索记录和词句消岐 |
| GSpell | 拼写建议工具，使用多个词相似算法，为拼写错误提供正确拼写形式 |
| BagOWordsPlus | 使用GSpell的词相似性算法，执行基于短词层次的词相似性信息检索 |
| dTagger | 词性标注(Part of Speech，POS)工具 |
| Visual Tagging Tool (VIT) | 通过不同视觉风格显示被标注文本，便利人工标引工作 |

### 2.4 支持性软件工具

(1)UMLS的术语服务

UMLS术语服务(UMLS Terminology Services，UTS)①于2010年12月上线，取代了之前的UMLS知识源服务器(UMLS Knowledge Source Server，UMLSKS)。UTS通过基于浏览器和Web服务客户端提供对UMLS知识源的浏览、查询和数据获取，主要工具包括：超级叙词表浏览器(见图2②)、语义网络浏览器和SNOMED CT浏览器，这些浏览器能够查询和获得UMLS的概念、语义类型、语义关系和SNOMED CT的内容[17]。

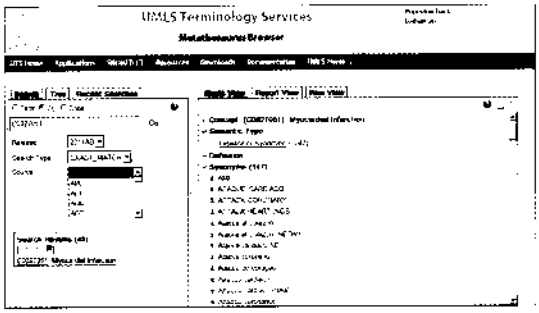


图2 UMLS术语服务的超级叙词表浏览器界面

(2)UMLS的安装和定制工具MetamorphoSys

MetamorphoSys是对UMLS进行本地安装和对UMLS进行定制化裁减的工具。用户可以通过该工具选择安装超级叙词表、语义网络、专家辞典中的一项或多项内容。当选择安装超级叙词表时，安装向导允许用户创建超级叙词表的子集，即可以选择去掉某些来源的数据文件，或者通过选项设置和过滤器进行定制和裁减，达到缩小容积、满足个性化要求的目的[17]。

(3)UMLS概念的文本映射工具MetaMap

MetaMap③是一个实现自由文本到UMLS概念映射的工具，即标记出生物医学文本中所含有的UMLS超级叙词表概念。MetaMap的应用非常广泛，如Medline数据检索，有研究表明[18]，它能够提高Medline文献信息检索的效果；同时，MetaMap在数据挖掘领域也有广泛的应用，包括临床发现、发现文献中的药物与疾病关系等；此外，MetaMap也是NLM自动标引系统的实现基础，用于为半自动和全自动标引生成推荐术语[19,20]。

① https://uts.nlm.nih.gov
② https://uts.nlm.nih.gov
③ http://mmtx.nlm.nih.gov

(4)语义表达工具 SemRep

SemRep[①] 应用自然语言处理技术和 UMLS 的专家辞典工具,将生物医学文本进行语句切分和词性标注,对所获得的术语应用 MetaMap 映射,获得其在 UMSL 超级词表中的相应概念,以及概念在语义网络中对应的语义类型和语义关系,并通过概念共现获得文本信息的主要论点,即该文本主旨内容的主语 - 谓词 - 对象形式的语义表达[21,22]。

## 3 UMLS 在智能信息检索中的应用

### 3.1 扩展查询

扩展查询(Query Expansion)是知识组织系统在信息检索中常见的应用形式,通过同义词、异形词映射,树状结构扩展和词组切分等,在原始输入的检索词基础上增加词形、语种和词义方面的扩展词,从而增加输入词匹配命中的可能性,也是利用同义词解决多词表不匹配问题的技术方法[23]。通过不同语种的同义词映射和扩展,能够实现跨语言检索。例如:中国生物医学文献数据库(CBM)、中国生物医学文献服务系统(SinoMed)提供基于 CUMLS 的主题词跨中、英文检索[24]。

UMLS 的超级叙词表为扩展检索提供了良好的基础。文献[25]利用 MetaMap 程序获取 UMLS 超级叙词表的相关概念进行扩展的方法,经过与基于检索反馈的扩展检索方法相对比,认为基于 MetaMap 的扩展方法与基于检索反馈的扩展检索相结合,其效果最优。事实上,由于医学信息领域大量文献使用 MeSH(Medical Subject Headings)进行标引,因此,在实际应用中,基于 UMLS 的扩展检索往往结合 MeSH 标引来实现,如 PubMed。

(1)PubMed 基于词映射的扩展检索

PubMed 通过自动术语映射(Automatic Term Mapping, ATM)[26] 机制来实现扩展检索。PubMed 根据用户输入的检索词,依次查找以下翻译表:MeSH 翻译表(MeSH Translation Table),期刊翻译表(Journals Translation Table),作者翻译表(Full Author Translation Table)和作者索引(Author Index)。其中,MeSH 翻译表包括以下内容:MeSH 的主题词(MeSH Heading),副主题词(Subheading),出版类型(Publication Type),MeSH 词的入口词(Entry Terms)即同义词,UMLS 是其中同义词的一个来源,可映射 UMLS 的概念、补充概念和补充概念的同义词[27]。

如果用户输入词与 MeSH 翻译表中上述词匹配,那么这些词将会被映射为适当的 MeSH 词而进行检索,检索者所使用的词和所映射的 MeSH 词将会分别作为文本词和 MeSH 词在文献的题名/文摘和 MeSH 标引字段进行检索。例如,用户的输入词为"Rash",通过翻译映射表,得到 MeSH 词标题"Exanthema",则检索式为("exanthema"[TIAB] NOT Medline[SB]) OR "exanthema"[MeSH Terms] OR rash[Text Word][26]。

此外,PubMed 还会对命中的主题词进行树状结构扩展,即经过词表翻译,得到 MeSH 的主题词后,自动搜索该主题及其树状结构中的下级子主题词作为检索词。例如,用户输入"feet"一词,PubMed 会将其翻译为"foot"[MeSH Terms] OR "foot"[All Fields] OR "feet"[All Fields],也就是说 feet 是 MeSH 词 foot 的入口词。具体匹配过程是在所有字段中检索该词组和组成该词组的各个单词,除非这些词映射为物质名称或标题词中包括独立数字或字母。例如"muscle atrophy",在 PubMed 中的检索式为"muscular atrophy"[MeSH Terms] OR ("muscular"[All Fields] AND "atrophy"[All Fields]) 或者 "muscular atrophy"[All Fields] OR ("muscle"[All Fields] AND "atrophy"[All Fields]) OR "muscle atrophy"[All Fields]。而检索词"protein c",则检索为"protein c"[MeSH Terms] OR "protein c"[All Fields],不作词组拆分[27]。

可见,PubMed 使用了基于概念的同义词扩展、等级结构扩展和词组切分扩展方法。

(2)扩展检索的效果评估

Hersh 等[28]利用 Medline 进行实验,结果表明使用 UMLS 进行扩展检索并不一定可以取得更好的检索效果。文献[23]介绍了使用 MeSH 和使用 UMLS 进行扩展查询的对比实验,结果显示基于 UMLS 的扩展查询 MAP 值,低于基准检索(即不使用扩展的检索)和基于 MeSH 的扩展检索。

实验[29] 比对了三个查询:基准查询,基于 MeSH 的扩展查询和基于 UMLS 的扩展查询集,每个查询分别对应三个不同的数据集:图表标题和题名(CT),图表标题、题名和段落文本(CTS),图表标题、题名和全文

① http://skr.nlm.nih.gov

(CTA)。扩展方法是利用 UMLS 的 MetaMap 程序获得映射词,为了避免使用扩展所带来的查询词过多导致与 MeSH 不相匹配的问题,在 MetaMap 的使用中,限制了语义类型,即在以下范围中查找扩展词:bpoc (Body Part,Organ,or Organ Componet), diap(Diagnostic Procedure), dsyn (Disease or Syndrome) 和 neop (Neoplastic Process)。三种查询在不同数据集的查询结果显示,利用 MeSH 在最小的数据集(CT)中,取得的检索效果最好,使用基于 UMLS 的效果在三个数据集中效果最差。

文献[30]提出,影响基于 UMLS 扩展检索实验结果主要有 4 个因素,即索引策略、检索系统、测试集、所提交的查询式。其中,最核心的问题是索引策略的构建。目前基于 UMLS 超级叙词表的扩展查询,主要有 4 种策略:SC,概念扩展的字符索引;ST,术语扩展的字符索引;WC,概念扩展的词索引; WT,术语扩展的词索引。通过对 MedlinePlus 数据集进行检索,4 种策略的对比检索结果显示:使用术语层扩展比使用概念扩展的准确率更高;基于字符的索引策略比基于词索引的索引策略的准确度高;使用术语扩展的字符索引准确度最高[30]。

### 3.2 语义检索

(1)基于语义关系的检索

语义检索与基于关键词逻辑匹配的传统检索方式相比,是基于概念和概念之间关系的模式匹配过程。在概念匹配的基础上,引入语义关系及关系匹配,是语义检索的重要特征[31]。

文献[32]和文献[33]探索了引入 UMLS 的语义关系应用于信息检索的两个关键过程——关系扩展查询和命中文献组织方面的研究。

①关系及关系扩展查询

关系扩展查询的基本思想是在识别和确定概念的基础上,通过 UMLS 语义网络,确定概念之间的关系,增加到原本的概念匹配过程中,形成概念 - 关系匹配。具体实现流程如图 3 所示:

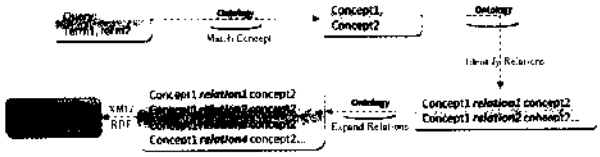


图 3 基于 UMLS 的关系查询[33]

1)用户输入术语,通过 UMLS 术语匹配,确定相应概念;

2)用户需求的概念表达,通过 UMLS 语义网络,获得相应的语义类型;

3)语义类型通过 UMLS 语义网络,获得语义关系作为候选关系待检索;

4)候选关系的同义和上下位关系被确定为候选概念族信息。

文献[32]给出一个实例,用户输入"liver cancer,food"这样的查询,通过 UMLS 的语义类型确定和语义关系查找,得到"cause"和"affect"两个语义关系,而"affect"关系有一个下位语义关系 treat,则以 XML 表达的语义关系查询式[32]如下所示:

```
<Query>
  <Concept_1>
    <Category> neoplastic process </Category>
    liver cancer
  </Concept_1>
  <Candidate Relations>
    <Concept_2_Concept_1>
      <Relation 1> cause </Relation 1>
      <Relation 2> affect
        <Hyponym> treat </Hyponym>
      </Relation 2>
    </Concept_2_Concept_1>
  </Candidate Relations>
  <Concept_2>
    <Category> food </Category>
    food
  </Concept_2>
</Query>
```

该检索式匹配目标文献时,命中候选句至少同时含有两个候选概念和至少一个候选关系。

②基于关系的文献组织

检索结果基于候选关系和相关关系进行组织。检索命中文献集被分为"cause"集和"affect"集,这两个结果集中的文献根据 UMLS 的关系结构,被进一步划分,即 treat 集被安排到 affect 集下,这与本体中的关系结构是一致的。同类的文献不必共享相似的词分布,但是有相同的概念关系[32]。

(2)检索结果的语义化表达

①Semantic Medline 的可视化语义表达

Semantic Medline① 是美国国立医学图书馆的语义 Web 应用原型系统,它对 PubMed 检索命中所返回的 Medline 数据进行自动文摘,应用 UMLS 和自然语言处理技术分析题名

① http://skr3.nlm.nih.gov/SemMedDemo/

和文摘中的主要含义，并将这种主要含义以主语－谓词－对象的三元组形式进行表达，通过可视化方式展现三元组关联网络，同时在图形中提供指向 PubMed、UMLS 等相关数据源的链接用于导航，并允许用户基于图形化的结果集进行浏览和二次检索[34]。

Semantic Medline 的主界面包括三个导卡页：Home、Search 和 Summarize。Search 页面调用 PubMed 检索 Medline 文献；Summarize 页面对命中文献进行自动文摘和主旨句的语义可视化展现。如图 4 所示，页面上端显示了通过 Search 页面检索和 Summarize 页面浏览所获得的当前检索 Session 的简要描述，包括检索词、数据源、日期和从 Medline 命中文献中提取出的论断（Predication）三元组数据量。所谓“论断”是一个以文本表达的关系论述的规范表达。在语义 Medline 中，论断是命中文献的文摘主旨，即从 Medline 的文本信息包括题名和文摘中提取的中心主题句[34]。

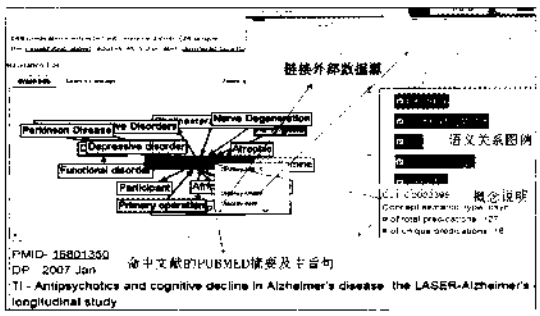


图 4 Semantic Medline 的检索结果语义表达

页面下方是可视化导卡，一个论断在图中表达为一个三元组，由两个参数和一个关系连结而成，分别用长方形和连线表示主语、对象和谓词。一个结果集中的相同论断被合并，并基于节点连接成网络。节点和连线根据含义进行加色。节点的颜色由 UMLS 的语义类型决定，如：Treatment of Disease，Substance Interactions，Diagnosis 和 Pharmacogenomics 等分别用不同的颜色表达。连线的颜色根据 UMLS 的语义关系而不同。连线即关系图例显示在右侧面板中，点击其中一个将显示或隐藏图中连线所代表的谓词，提供聚焦视图。点击图中的元素，右侧的信息导卡面板会显示相应的信息，包括谓词和论断在文献结果集中出现的次数、UMLS 概念标识和语义类型，以及与外部资源包括 UMLS Semantic Navigator，Entrez Gene，GHR 和 OMIM 等可能的链接；点击连线还可以看到当前主旨句的 Medline 文摘信息，包括 PubMed 标识符、题名和文摘，主旨句在引文句中高亮显示。

语义 Medline 作为一个原型系统，实际上提供了几种语义检索场景的支持：

1）检索结果集的语义化表达：语义 Medline 通过文献主旨句的提取、主旨句的语义表达、语义关联网络的构建，可帮助用户快速了解检索结果的主题含义及其语义关系，以进行文献选择和信息利用。

2）语义浏览和导航：通过点击可视化关系图中的节点和连线，可以实现语义浏览功能，具体包括：通过视图聚集，选择部分检索结果；连接外部数据源。

3）限制检索：在检索结果语义表达图中进一步输入概念，更加直观地进行限制检索和二次检索。

②Semantic Medline 的实现

Semantic Medline 是基于 Java 的三层 Web 应用，底层使用 MySQL 数据库存储 Semantic Medline 数据，包括从 UMLS 超级词表和 Entrez 基因数据库所提取的语义论断。数据库预先存储从文本文件转换而来的、含有 SemRep 的输出和使用 Perl 脚本的超级词表和 Entrez 基因数据[35]。

从实现上来看，语义 Medline 实现的核心是对 SemRep 和 MetaMap 应用程序的调用。SemRep 检测输入文本的每个句子，识别出语句含义中表达概念和关系的部分。如论断“Genes AFFECTS Circadian Rhythms”是从句子“Clock genes are the genes that control circadian rhythms in physiology and behavior”提取出的。概念“Genes”和“Rhythms”来自于 UMLS 的超级叙词表，谓词“AFFECTS”是语义网络中的一个关系语义，MetaMap 则在这一过程中实现文本与超级词表概念的映射，从而获得相应的语义类型并查找语义关系[35]。

### 3.3 问答式检索

问答式检索是自动问答系统（Question Answering System，QA）的形式之一[36]。实际上它是一种特定的信息检索方式，用户以自然语言提问式作为检索输入，系统通过文献检索，以命中文献中的特定文本作为回答结果输出。问答式检索的技术核心是用户提问的处理和文献信息的检索。在开放式领域中，问答系统的核心是命名实体的识别，如人名、地名、机构、时间等。而构建医学专业领域的问答系统，则需要识别与领域相关的实体名称，为了达到这个目标，University of Toronto 的 EpoCare 临床问题系统通过医学文本训练，识别出 UMLS 语义网络中的语义类型，作为医学领域命名实体标注的来源[2,37]。

文献[38]设计了引入 UMLS 的临床医学自动问答系统，比较完整地展示了问答式检索的基本流程和对 UMLS 的深入应用，类似的工作参见文献[39]。问答式检索系统的主要流程如图 5 所示。

（1）问题分析：问题分析的作用是将用户提问分析、输出为相关的概念和语义关系，该过程使用

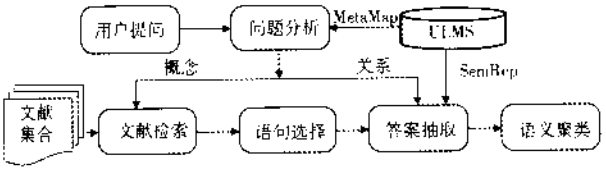


图 5 基于 UMLS 的问答式系统流程[38,39]

MetaMap 工具，将所析出的医学概念传递给后面的“文献检索”组件，将语义关系传递给“答案抽取”组件。同时，该过程也会对提问的类型进行划分和确定，生成三元组式的陈述，以用于匹配后面 SemRep 生成的表达。

(2)文献检索：使用开放源码引擎 Lucene，对接收到的概念，利用 MetaMap 进行概念扩展后再进行检索，选择命中结果权重最高的前 20 个文献进一步处理。

(3)语句选择：在命中文献中进一步缩小范围，选择相关语句。选择的方法是将用户提问中的医学概念、同义词，使用“and”操作符连接这些术语并进行检索，命中得到相关的候选句。

(4)答案抽取：在问题分析阶段，已经得到问题的语义表达，在句子选择阶段，得到有可能对问题回答有帮助的相关句子。引入 SemRep 来识别候选句子中的关系，从而获得候选句子的语义表达，将其与用户提问的语义表达相匹配，从相关句子中生成短词水平的回答。

(5)语义聚类：如果问题答案为某一医学概念，则基于 UMLS 的树状关系结构进行语义聚类，以该概念的上位概念或语义类型作为答案分类。

该实验在多个环节利用了 UMLS 的两个语义工具 SemRep 和 MetMap，从实验结果来看，对于事实型问题，该问答系统表现非常出色，而对于含有约束条件的复杂问题，则查准率有所下降。

## 4 结 语

UMLS 对智能检索的支持能力主要表现在以下两个方面：

(1)规模空前的概念体系是 UMLS 坚实的基础，而其特色和优势在于不仅完整保留了概念之间的各种关系，而且进一步完成梳理和重构，通过语义网络在更高层次、更有效地对语义关系进行了规范和控制；

(2)UMLS 工具化的特征非常显著，灵活多样的工具促进了 UMLS 知识源能够得到方方面面的使用，而各种方法、手段、算法的广泛复用甚至是工具调用工具形成新的工具(如 SemRep 工具本身调用了专家辞典工具和 MetaMap)，促进了工具本身的应用推广和优化，不断推动 UMLS 的应用和建设。

对于集成式知识组织系统在智能检索中的应用，需要深入思考和解决以下问题：

(1)综合性、集成式词表系统的优势与应用效果

UMLS 的官方网站上明确表示，UMLS 作为综合多领域、多词表的集成系统，不能保证它在某个领域或系统中的应用是最优的，而主要目标是起到“联通”作用，即联通不同词表在不同领域和系统中的应用。本文所调研的扩展检索效果评估也证明了这一点，应用 UMLS 的扩展检索效果低于基于 MeSH、甚至是不使用扩展检索的效果。因此，对于集成式词系统，如何设计应用才能最大限度地发挥其作用，是一个值得深入思考、研究和不断实验、评估的问题。

(2)扩展检索的效果与结果集收敛

扩展检索中的词形扩展、语种扩展以及等级结构扩展和相关词扩展等，会带来检索结果集的增大，有可能提高查全率，但随之而来是否会带来误检率、漏检率和特异率[40]的增长，还需要实验加以评测；同时，结果集的增大会在信息的筛选和过滤方面给用户带来更多的困扰，因此，扩展检索结果集如何进行收敛，需要结合用户的检索行为和认识习惯在设计层面进行综合思考，例如一方面可以由用户进行选择是否扩展或者明确提示检索词和扩展词，另一方面通过结果集聚类、多维分面、图形化展示等手段对检索结果集进行有效的划分。

(3)关系检索的局限

从本文所调研的各种系统和原型来看，基于概念、特别是概念之间的语义关系进行内容表达和检索，已成为智能化信息检索的主要特征之一。有研究表明基于关系的信息检索技术优于基于项(Term)或基于语义(Concept)的检索技术[41]。但实现关系检索需具备一定的前提条件，如关系的提取和确定、关系的取值和属性等，如何借助类似于 UMLS 的语义网络和 WordNet 等的语义关系控制手段，在大规模数据环境中有效地解决上述问题，是实现关系检索的关键。

## 参考文献：


[1] 의료정보학세미나. UMLS Applications[EB/OL]. [2012-02-20]. http://www.snu-dhpm.ac.kr/pds/files/UMLS%20Applications_%ED%95%9C%EC%8A%B9%EB%B9%88.pdf.

[2] U.S. National Library of Medicine. UMLS Applications[EB/OL]. [2012-02-20]. http://www.nlm.nih.gov/research/umls/implementation_resources/applications.html.

[3] Fung K W, Hole W T, Srinivasan S. Who is Using the UMLS and How-Insights from the UMLS User Annual Reports[C]. In: *Proceedings of the* 2006 *AMIA Annual Symposium*. American Medical Informatics Association, 2006: 274-278.

[4] 维基百科. 一体化医学语言系统[EB/OL]. [2012-02-20]. http://zh.wikipedia.org/wiki/一体化医学语言系统. (Wikipedia. Unified Medical Language System[EB/OL]. [2012-02-20]. http://zh.wikipedia.org/wiki/一体化医学语言系统.)

[5] U.S. National Library of Medicine. Fact Sheet Unified Medical Language System® [EB/OL]. [2012-02-20]. http://www.nlm.nih.gov/pubs/factsheets/umls.html.

[6] 方平. 试论一体化医学语言系统(UMLS)超级叙词表的特点[J]. 图书情报工作, 1998, 42(10): 26-29, 41. (Fang Ping. The Characteristics of the UMLS Metathesaurus[J]. *Library and Information Service*, 1998, 42(10): 26-29, 41.)

[7] 2011AB UMLS® Release Notes and Bugs[EB/OL]. [2012-02-20]. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/notes.html.

[8] Lindberg D A, Humphreys B L, McCray A T. The Unified Medical Language System[EB/OL]. [2012-02-20]. http://www.schattauer.de/en/magazine/subject-areas/journals-a-z/methods/contents/archive/issue/1198/manuscript/14376/download.html.

[9] Tilley C, Willis J. The Unified Medical Language System What is it and How to Use It? [OL]. [2012-02-20]. http://www.cs.rutgers.edu/~mdstone/class/336/umls.pdf.

[10] Kleinsorge R, Willis J, Browne A. UMLS Overview[OL]. [2012-02-20]. http://www.nlm.nih.gov/research/umls/pdf/AMIA_T12_2006_UMLS.pdf.

[11] Geissbuhler A, Miller R A. Clinical Application of the UMLS in a Computerized Order Entry and Decision-Support System[OL]. [2012-02-20]. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2232318/pdf/procamiasymp00005-0356.pdf.

[12] 方平. 试论一体化医学语言系统语义网络的结构与特点[J]. 情报学报, 1999, 18(2): 129-134. (Fang Ping. Study on Characteristics and Structure of Semantic Network of UMLS Knowledge Sources[J]. *Journal of the China Society for Scientific and Technical Information*, 1999, 18(2): 129-134.)

[13] Current Semantic Types[EB/OL]. [2012-02-20]. http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html.

[14] 朱彦慧, 腾吉斯. 一体化医学语言系统及其对我国的借鉴作用[J]. 中国科技术语, 2010, 12(4): 15-18. (Zhu Yanhui, Teng Jisi. The UMLS and Its Reference to Standardize Chinese Medical Terminologies[J]. *China Terminology*, 2010, 12(4): 15-18.)

[15] 维基百科. 专家辞典[EB/OL]. [2012-02-20]. http://zh.wikipedia.org/wiki/UMLS#.E4.B8.93.E5.AE.B6.E8.BE.9E.E5.85.B8. (Wikipedia. SPECIALIST Lexicon[EB/OL]. [2012-02-20]. http://zh.wikipedia.org/wiki/UMLS#.E4.B8.93.E5.AE.B6.E8.BE.9E.E5.85.B8.)

[16] The SPECIALIST NLP Tools[EB/OL]. [2012-02-20]. http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html.

[17] UMLS® Reference Manual[EB/OL]. [2012-02-20]. http://www.ncbi.nlm.nih.gov/books/NBK9676/.

[18] Aronson A R. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program[OL]. [2012-02-20]. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243666/?page=4.

[19] Aronson A R, Lang F M. An Overview of MetaMap: Historical Perspective and Recent Advances[OL]. [2012-02-20]. http://www.lhncbc.nlm.nih.gov/lhc/docs/published/2010/pub2010033.pdf.

[20] 张云秋, 冷伏海. MetaMap的文本映射原理及其对检索效果影响的研究[J]. 情报学报, 2007, 26(3): 345-349. (Zhang Yunqiu, Leng Fuhai. Study on the Principle of Text Mapping and Its Effect on Information Retrieval of MetaMap[J]. *Journal of the China Society for Scientific and Technical Information*, 2007, 26(3): 345-349.)

[21] Rindflesch T C, Aronson A R. Semantic Processing for Enhanced Access to Biomedical Knowledge[OL]. [2012-02-20]. http://skr.nlm.nih.gov/papers/references/semwebapp.5a.pdf.

[22] Rindflesch T C, Fiszman M, Libbus B. Semantic Interpretation for the Biomedical Research Literature[EB/OL]. [2012-02-20]. http://ai.arizona.edu/mis596a/book_chapters/medinfo/Chapter_14.pdf.

[23] DíazGaliano M C, GarcíaCumbreras M A, MartínValdivia M T, et al. Query Expansion on Medical Image Retrieval: MeSH vs. UMLS[C]. In: *Proceedings of the 9th Cross-Language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access*. 2009.

[24] 李丹亚, 胡铁军, 李军莲, 等. 中文一体化医学语言系统的构建与应用[J]. 情报杂志, 2011, 30(2): 147-151. (Li Danya, Hu Tiejun, Li Junlian, et al. Construction and Application of the Chinese Unified Medical Language System[J]. *Journal of Intelligence*, 2011, 30(2): 147-151.)

[25] Aronson A R, Rindflesch T C. Query Expansion Using the UMLS® Metathesaurus® [OL]. [2012 - 02 - 20]. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233565/pdf/procamiaafs00001-0521.pdf.

[26] Carlin B G. PubMed Automatic Term Mapping[J]. *Journal of the Medical Library Association*, 2004, 92(2): 168.

[27] NLM® Training: PubMed® [EB/OL]. [2012 - 02 - 20]. http://www.lib.hiroshima-u.ac.jp/online_tu/dl/pm_workbook.pdf.

[28] Hersh W, Price S, Donohoe L. Assessing Thesaurus - based Query Expansion Using the UMLS Metathesaurus[OL]. [2012 - 02 - 20]. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244120/pdf/procamiasymp00003-0379.pdf.

[29] Galiano D, Cumbreras G, Valdivia M. SINAI at ImageCLEFmed 2008 [OL]. [2012 - 02 - 20]. http://clef.isti.cnr.it/2008/working_notes/diaz-paperCLEF2008.pdf.

[30] Lu K, Mu X M. Query Expansion Using UMLS Tools for Health Information Retrieval [EB/OL]. [2012 - 02 - 20]. http://www.asis.org/Conferences/AM09/open-proceedings/papers/12.xml.

[31] 张晓林. Semantic Web 与基于语义的网络信息检索[J]. 情报学报, 2002, 21(4): 413 - 420. (Zhang Xiaolin. Semantic Web and Semantic - based Networked Information Retrieval[J]. *Journal of the China Society for Scientific and Technical Information*, 2002, 21(4): 413 - 420.)

[32] Chen M. Exploring the Use of Ontological Relations in Information Retrieval[OL]. [2012 - 02 - 20]. http://ischools.org/images/iConferences/iConference2009_poster_Final.pdf.

[33] Chen M. Using Ontological Relations to Enrich Query Semantics [OL]. [2012 - 02 - 20]. http://mail.asis.org/Conferences/AM09/posters/96.pdf.

[34] Semantic Medline Help[EB/OL]. [2012 - 02 - 20]. http://skr3.nlm.nih.gov/SemMedDemo/jsp/help.jsp.

[35] Kilicoglu H, Fiszman M, Rodriguez A, et al. Semantic MEDLINE: A Web Application for Managing the Results of PubMed Searches [OL]. [2012 - 02 - 20]. http://pathema.jcvi.org/Pathema/presentations/kilicoglu2008.pdf.

[36] 许德山, 乔晓东, 朱礼军. 问答系统分类研究及新进展[C]. 见: 第二十一届全国计算机信息管理学术研讨会论文集, 2007. (Xu Deshan, Qiao Xiaodong, Zhu Lijun. Classification and New Progress of Question - Answering System[C]. In: *Proceedings of the 21st National Conference on Computer Information Management*, 2007.)

[37] Delbecque T, Jacquemart P, Zweigenbaum P. Indexing UMLS Semantic Types for Medical Question - Answering[OL]. [2012 - 02 - 20]. http://cluster.cis.drexel.edu:8080/sofia/resources/QA.Data/PDF/M_2005_ENMI_Delbecque_Indexing_UMLS_Semantic_Types_for_Medical_Question-Answering-2817493248/M_2005_ENMI_Delbecque_Indexing_UMLS_Semantic_Types_for_Medical_Question-Answering.pdf.

[38] Wang W M, Hu D W, Feng M, et al. Automatic Clinical Question Answering Based on UMLS Relations[OL]. [2012 - 02 - 20]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.4658&rep=rep1&type=pdf.

[39] Terol R M, Martínez - Barco P, Palomar M. A Knowledge Based Method for the Medical Question Answering Problem[OL]. [2012 - 02 - 20]. http://rua.ua.es/dspace/bitstream/10045/4511/3/manuscriptR1.pdf.

[40] 李毅, 庞景安. 基于多层次概念语义网络结构的中文医学信息语义标引体系和语义检索模型研究[J]. 情报学报, 2003, 22(4): 403 - 411. (Li Yi, Pang Jingan. Research on Semantic Indexing System and Semantic Retrieval Model for Chinese Medical Information Based on Multilayer Conceptual Semantic Network Structure[J]. *Journal of the China Society for Scientific and Technical Information*, 2003, 22(4): 403 - 411.)

[41] 李岩, 文健, 李舟军. 一种改进的基于关系的信息检索技术[J]. 计算机科学, 2008, 35(7): 145 - 150. (Li Yan, Wen Jian, Li Whoujun. Improved Relation - based Information Retrieval Technology[J]. *Computer Science*, 2008, 35(7): 145 - 150.)

(作者 E - mail: bhy@istic.ac.cn)