科技文献检索系统语义丰富化框架的设计与实践

谢靖[1] 王敬东[1] 吴振新[1] 张智雄[1] 王颖[1] 叶志飞[1]
[1]（中国科学院文献情报中心　北京　100190）

摘要：[目的/意义]本文期望通过采用数据挖掘、语义识别、知识关系计算等技术方法来提升科技文献检索系统的服务功能和效果，使之能够呈现更加丰富的知识化语义信息，将更多的知识点和知识关系展现给用户。[方法/过程] 本文应用 semrap 和 clausIE 数据挖掘和关系计算工具，识别和抽取科技文献中的语义对象，分析、计算、构建语义关系，并将得到的语义对象和语义关系设计建立多维语义索引树，设计了新的数据组织呈现模型。[结果/结论] 研发语义丰富化检索示范系统，在科技文献检索系统中充分揭示语义信息，给用户带来更多的知识内容层面的导航、关联、发掘和揭示，同时分析了设计模型的优势与不足。

关键词：语义丰富化　语义知识组织　多维索引　语义关系呈现
分类号：TP391

Research and Practice on Semantic Enrichment of Scientific Literature Retrieval System

Xie Jing[1]　Wang Jingdong[1]　Wu Zhenxin[1]　Zhang Zhixiong[1]　Wang Ying[1]　Ye Zhifei[1]
National Science Library, CAS　Beijing　100190

Abstract: [Objective] This paper aims to enhance the scientific literature retrieval system, which can present more abundant semantic information, and will show more knowledge and knowledge relations to users based on data mining, semantic recognition, knowledge organization technology. [Methods/Process] This paper recognizes knowledge objects in the scientific literature, calculates and extracts semantic relations among the objects using the semrap and clausIE tools. It builds semantic multidimensional index on the semantic objects and semantic relations, presents a new data organization model design. [Results/Conclusions] This semantic enrichment retrieval demonstration system built in this paper can fully reveals the semantic information; can brings more knowledge level navigation, association, excavation and disclosure to users.

Keywords: Semantic Enrichment, Knowledge Organization, Multidimensional Index, Semantic Relation Presentation

引言

随着数据挖掘、语义识别、知识关系计算技术的不断发展和在科技文献中的应用，人们更希望在科技文献的检索系统中，能够呈现更加丰富的语义化内容，将更多的知识点和知识关系展现给用户。语义丰富化框架的设计目标是将多种类型的语义知识对象，语义对象之间丰富的关联关系等知识化信息在科技文献检索系统揭示出来，在数据挖掘、语义识别技术的基础上，改变现有单一关键词导向的检索系统，重新组织语义丰富化的数据，以呈现语义知识与语义

关联信息。

本文在语义丰富化框架设计试验的过程中，选择来自于 PubMed[1]的医学领域的 Migraine Disorder、Heart Diseases 这 2 个主题近 2 年内的文章集合作为示范系统的试验数据集，采用数据挖掘计算较为成熟的 semrap 和 clausIE 作为基础数据挖掘分析工具，并设计多维语义数据组织索引模型，研发了检索示范系统以探索科技文献检索的语义丰富化。

1 语义丰富化总体框架设计

图 1 语义丰富化总体设计框架图

如图一所示，科技文献检索系统语义丰富化主要工作分为两个部分：

（1）语义标引：主要对文献中出现的知识对象做深度标引。首先标引出文献的关键词，并识别出知识对象所属的类型（即它是什么）；其次计算出来各个知识对象之间的关联关系。语义标引主要工作包括：文献线索的标引、内容语义标引、语义关系抽取、句法关系抽取。

- 文献线索的标引：不仅包括文献元数据标引，如作者、出版社、出版年等描述数据。还需要根据文献提供的摘要文本数据，将文献线索切分为有独立内容含义的句子和段落，如计算切分文献中研究目的、研究方法、研究工具、研究结果等。
- 内容语义标引：实现学术论文中的问题、理论、方法、技术手段、工具、模型、结论等内容的标引。
- 语义关系抽取：根据在同一个句子中计算标引得到的知识对象，根据 UMLS 和 STKOS 知识组织体系[2]，查询每两个知识对象可能存在的语义关系[3]，并将发现的这些语义关系记录为 SPO 三元组[4]。
- 句法关系抽取：根据句法关系[5]计算，将长句拆分成短句和子句，并在这些子句中识别出来主谓宾关系，将主谓宾关系以 SPO 三元组的方式记录。一个句子可能拆分为多个 SPO 三元组。

（2）语义索引：将标引抽取后得到的内容，根据不同维度的数据，构建多维度的语义索引体系，将语义知识有机组织起来，便于语义检索平台使用。

- 文献索引：文献索引层是对文章的标题、作者、发表时间等元数据描述进行索引；句子和段落层索引是将文摘切分成段落和句子后，对句子和段落索引。

- 知识对象索引：从文本中识别标引得到的术语和实体，本文中统一称为知识对象[6]。根据语义丰富化示范系统的设计要求，知识对象索引包括知识对象索引和知识对象属性索引两个部分。知识对象索引用于将用户输入关键词识别为语义对象，实现对输入关键词规范作用。知识对象属性索引用于对知识对象的各项属性进行检索查询和分类展示。
- 知识对象关系索引：语义标引工作计算得到语义关系和句法关系，文本将这两种关系合并后统称为知识对象关系。它们均以 SPO 三元组的方式表达，并构建三元组索引，从而实现对知识对象语义关系的检索查询和关联关系揭示。

2 语义标引的功能设计与实现

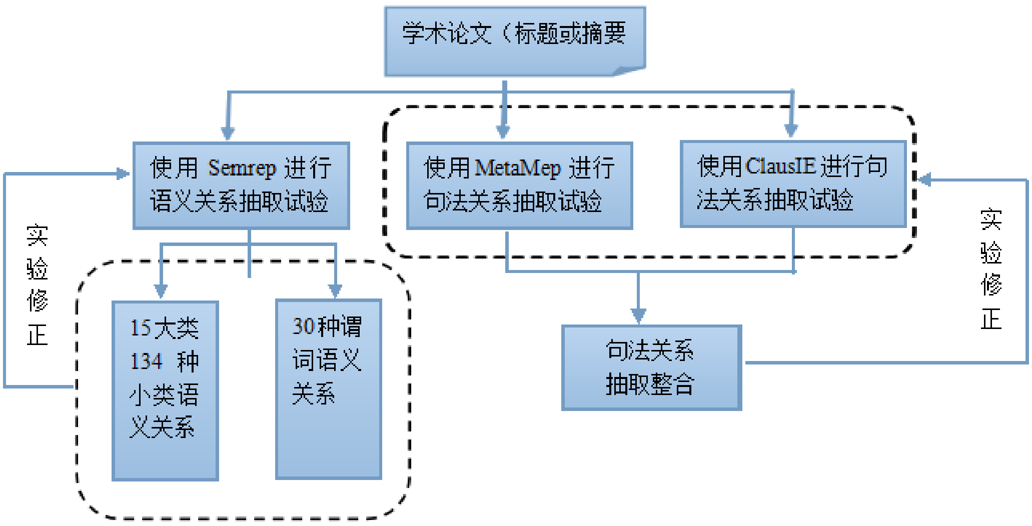

图 2 语义标引流程图

对选取文献数据的标题和摘要进行语义标引，参照 UMLS、STKOS 将医学领域语义对象划分 15 个大类、134 个小类为使用 Semrap 和 MetaMap 工具对文献中的重要语义对象进行标引抽取。使用 MetaMap 和 ClauseIE 实现对语义关系的计算识别。工作流程如图 2 所示，左边流程代表文本内容通过 Semrap 语义对象标引，标引得到实验数据参照选取的 15 个大类、134 个小类进行规范映射等实验数据修正。标引工作内容包括：

（1）实现深入文献内容的重要对象标引，基于海量科技文献（20 万条）、STKOS 的本体和超级词表（UMLS），通过 Semrap 和 MetaMap 工具实现 15 个大类、134 个小类的重要内容和重要对象的标引抽取。

（2）实现知识对象标引，标引规范后知识对象 4935 条（未规范对象 20 万

条）

2.2 数据语义化组织与规范

如图 2，右边流程代表对对象关系的计算、组织和规范，MetaMap 工具实现对规范语义关系的计算识别，将语义对象关系识别为 30 个规范关系。ClauseIE 工具实现对句法树关系的识别。将 MetaMap 和 ClauseIE 两种工具识别的语义关系数据合并整合，参照 MetaMap 选取的 30 个规范关系对实验数据规范修正。完成的数据组织、规范工作包括：

（1）实现文献内部知识对象的语义关系标引，通过 Semrap 和 MetaMap 工具实现科技文献中 30 种语义关系、段落关系的抽取，挖掘知识对象之间潜在的语义关系。

（2）实现文献内容的句法关系标引，通过 ClauseIE 工具实现科技文献中句法关系（SPO）抽取，发现知识对象（关键词、术语）之间潜在关联关系。

（3）整合语义关系和句法关系标引，通过试验 1116 篇文献摘要中提取的 SPO 关系 50204 条，其中语义关系 41590，语法关系 8614 条。

2.3 关键问题解决方案

（1）标引内容与 Mesh 词表映射

Semrep 处理后的结果如下：

SE|00000000||tx|1|entity|C0006142|Malignant neoplasm of breast|neop|Breast cancer|1000|1|13

含义如下：

Semrep 标签	数据字段含义描述
SE	Semrep 工具标记
0000	文章标识
tx	文本来源标记
1	实体在文本中的位置
Entity	术语类型
C0004142	Mesh 词表中术语代码
Malignant neoplasm of breast	Mesh 种的标准术语
neop	语义关系缩写
Breast cancer	文本中出现的词
1000	置信度

表 1 Semrep 语义标引字段描述表

红色字段（例子中的 neop）为 134 种小类的语义关系缩写，目前本文已经收集了 Mesh 词表对应的 134 种小类 15 个大类的英文全称、英文缩写及中文名称。通过红色字段进行关联，建立起文本识别术语与 Mesh 词表映射关系。从而解决了 Semrep 处理后的结果与 15 大类和 134 种小类对应关系问题。

（2）ClausIE 抽取出的主语（S）、谓词（P）与 UMLS 超级词表的对应

ClausIE 是按照句法关系来抽取三元组，所以和 Semrep 种抽取的实体不能完全匹配；同时 Semrep 只能抽取语义动词，其他动词都被忽略掉。对于前一种情况考虑能否通过模糊匹配的方式保证实体的对应；对于第二种情况，考虑从 MetaMap 中提取出动词，然后进行匹配。保障试验数据的规范性和一致性。

3 语义索引的功能设计与实现

图 3 语义索引架构设计图

3.1 语义索引的基本功能

语义索引设计目标是揭示语义对象和对象间多种语义关系，改变了当前单一维度索引的方式，使用多颗索引树整合协同工作，从多维度呈现语义内容。如图 3 所示，语义索引以知识对象为核心，遵循用户使用流程，从检索关键词出发，通过知识对象索引对输入关键词进行语义识别和语义消歧；然后通过知识对象关系索引，遍历知识网络，导航、筛选所需关联知识；通过桥接索引确定知识对象所在的句子、段落；最后通过文献索引查询、展示包含相关知识内容的文献信息。基于上述 4 个步骤，将索引分为 4 个功能部分：

（1）知识对象索引

- 知识对象索引：索引知识对象的全称、简称、别名等，将用户检索输入的关键词转换为相关知识对象，实现语义检索转变。
- 知识对象属性索引：检索并展示知识对象的各项属性，发现语义冲突的关键词，实现语义消歧功能。

（2）语义关系索引

- 知识对象语义关系索引：索引文本中出现的知识对象间的语义关系（语义关系是 UMSL 或 STKOS 规范后的关联关系），实现语义关系的检索和分析展示功能。
- 知识对象语法关系索引：索引文本中出现的知识对象间的语法关系（语法关系是 NLP 句法分析得到未规范关联关系），用于区别语义和语法关系的检索和分析展示。

（3）桥接索引

- 对象-文献关系索引：实现知识对象和存在文献位置的映射，并同时用于分析和揭示语义知识对象的共现关系。

- 对象-段落关系索引：实现知识对象和存在段落位置的映射，并同时用于分析和揭示语义知识对象的共段关系。
- 对象-句子关系索引：实现知识对象和存在句子位置的映射，并同时用于分析和揭示语义知识对象的共句关系。

（4）文献索引

- 元数据索引：索引文献的元数据描述信息，包括标题、作者、出版年等元数据，用于文献基础信息的展示。
- 文献内容索引：对文章摘要（或全文）的内容索引，用于文献内容展示和相关知识对象和知识关系的高亮显示等功能。

本文试验共实现索引文献 1116 篇，段落 4023 个，句子 7684 个，索引规范知识对象 4935 条。索引知识关系 50204 条。

3.2 关键问题及解决方案

（1）输入的关键词与规范知识对象的映射

试验中可能出现输入关键词语与索引知识对象不能完全匹配问题，无法映射到准确的规范知识对象；输入的一个关键词可能包含多种含义，发生语义识别歧义，无法明确映射到具体知识对象。

对第一问题，本文采用索引模糊匹配方法，选取匹配分值最高的知识对象，并列出匹配的前 5 条列表通知用户，以便用户再次修正实现语义识别。第二个问题，则将给用户列出不同含义的知识对象，由用户选择实现语义消歧。后期工作可以考虑使用用户行为上下文进行智能语义消歧。

（2）知识对象关联关系统计揭示

知识对象关系都以三元组 S-P-O 的方式在 Apache Solr 建立索引，为了方便分析数据关系，三元组索引中加入冗余字段，即索引采用对主语（S）建立索引对（PO）分面，对宾语（O）建立索引对（SP）分面的方法。利用 Solr 的分面和频次统计功能，在检索知识对象时对（PO）和（SP）分面，即可统计揭示检索结果中出现频次最高的语法和语义关系，从而帮助用户发现潜在知识关联。

4 语义丰富化试验系统的数据组织

图 4 语义丰富化检索的数据组织结构图

为实现语义丰富化检索示范平台，系统将数据组织为 4 个维度，如图 4 所示，

第一维度是文摘层，对文章的标题、作者、发表时间等元数据表达揭示；第二维度是句子和段落层，将文章切分成段落和句子，对句子和段落表达揭示；第三维度是事实层，即对句子的语义化切分。用于知识对象计算得到语法关系和句法关系表达揭示；第四维度是知识对象层，对文本中识别出来的知识对象（术语和实体），以及知识对象属性表达揭示。

从下而上的视角看，第三、四维度将科技文献拆分为知识对象和知识对象的关联，从而形成了科技知识网络，用于语义化的查询与关联导航。第一、二维度结合文献的段落和句子，用于定位知识存在于科技文献的具体位置，便于用户详细关联阅读。

5 语义丰富化示范平台

语义丰富化示范平台围绕用户的知识化应用需求进行设计，用户检索流程一般为输入关键词，展示知识关系，关联导航深层具体知识点，查看知识所在具体文章。语义丰富化示范平台的研发实现了这 4 个功能：

（1）数字对象语义识别与检索功能：将用户输入关键词识别语义对象，使用语义进行检索。

（2）检索结果知识关系揭示功能：对检索内容，展示周边知识关系网络，揭示知识全貌。

（3）语义关系的关联导航：根据语义关系，关联导航深层具体知识点。通过语义关联导航，过滤更精确的检索结果。

（4）具体文章的语义化阅读：查看知识所在具体文章，对知识点和知识关联关系进行高亮显示辅助阅读。

功能介绍如下：

5.1 实现数字对象语义识别与检索

图 5 语义识别功能展示图

示范系统能够根据用户输入关键词识别出相应的语义对象，展示相关语义对象的解释。如图 5 所示：输入 headache 检索关键词，系统识别 headache 相关的语义对象，它是属于“体征或症状”的类型范畴。同时给出关于 Headache 的百科词条解释和相关的图片。

较之传统文献检索，这部分功能的优势在于可以规范用户输入，将模糊的关键词匹配检索转变为具有语义特征的语义对象检索，从而让语义丰富化检索更加精准。同时语义识别功能可以标示出语义对象的所属类型（或范畴），从而辅助用户进行语义消歧。避免传统关键词检索出现的语义偏差。

5.2 检索结果知识关系揭示

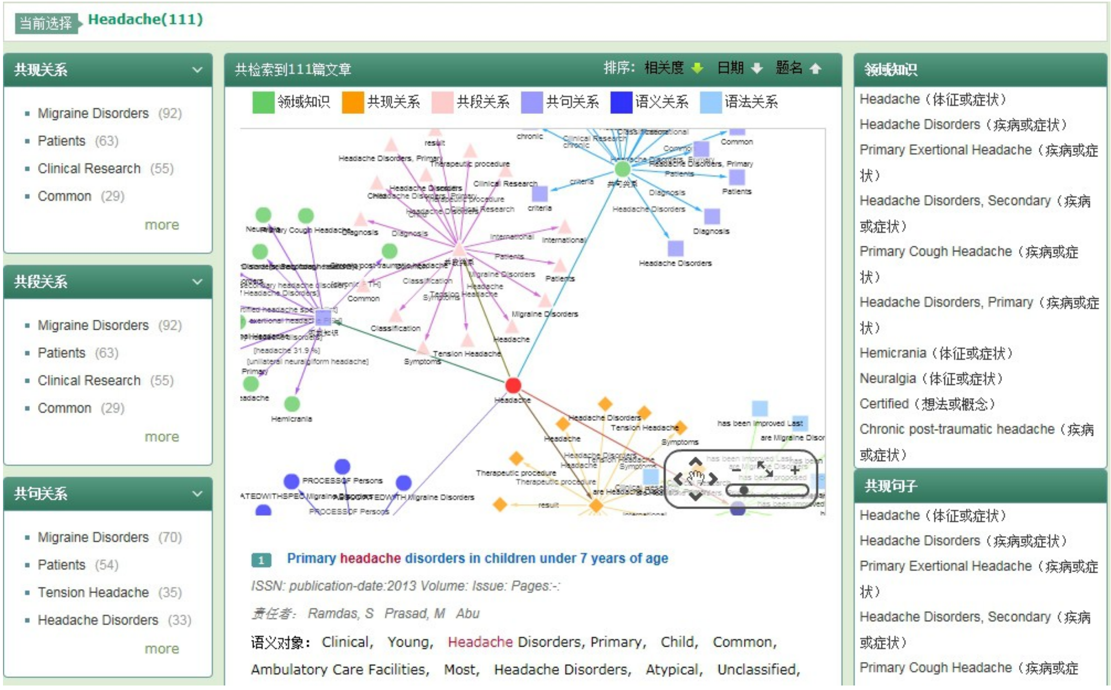

图 7 检索结果知识关系展示图

知识关系揭示功能在检索结果中，以图形方式揭示了涉及的知识对象、知识之间的语义关系、知识所在的文章片段（句子、段落等），如图 7 所示。这些知识及关联关系用以图形化的点和边的方式展示出来，使用不同的颜色的点代表不同类型的知识对象，用边表示知识之间的语义关系。可以通过点击-关联-导航的方式，让用户深入发现自己需要的知识。

示范系统能够清晰地展示出命中的知识关系以及这些知识存在的文章、句子段落。揭示重要句子和知识关系对科研人员判断该内容是否满足其检索需要有很大的帮助。本文认为使用相关知识对象、事实、句子、段落检索代替全文检索，对精准的语义知识发现更有帮助。用户可以点击语义化的知识对象、句子段落，关联链接查看文献全文。

5.3 实现语义关系的关联导航

图 6 语义关系导航展示图

语义关联导航功能根据检索输入匹配语义对象，在对检索结果文献中统计出共现、共段落、共句关系的语义对象。并实现关联语义对象的导航浏览，便于科研用户从潜在的关联语义对象中发现有价值的内容，并通过导航功能筛选出这些科技文献。如图 6 上半部分所示：查询 Headache 时共现关系、共句子关系、共段落关系出现 Migraine Disorders， Clinical Research 等，可能对科研人员起到启示的作用。

同样，SPO 语义关系和句法关系分面揭示，以谓词+宾语（知识对象）的分面统计方式揭示潜在语义、句法关系。如图 6 下半部分所示：检索 Headache，可以揭示发现儿童治疗(Process of Child)，治疗青春期(Process of Adolescent)等深层专业领域知识的文章，并可以揭示出相关治疗药物（followed eplepsy）的研究论文等，给科研人员提供明确的知识关系启发和导向。

示范系统通过语义关联、导航功能展现的知识共现关系及语义关系，是根据数据统计方式揭示出来的，有助于发现隐含的知识关联信息，也可帮助科研人员发现潜在的新知识关系，探索学科交叉领域的新研究点。扩展科研人员的研究思路，辅助科技创新。

5.4 单片文献的语义化辅助阅读

图 7 单片文献的语义化辅助阅读展示图

如图 7 所示，语义化辅助阅读功能在查看单片文献时，可以将计算识别的语义对象和知识之间关系高亮展示出来。如图 7 所示，左侧的树形列表展示的是该篇文献中计算识别的语义知识对象，将这些语义知识对象按照类型分到不同的组中，用不同颜色标示。中间主体部分是文献的文摘信息，选中某个类型的知识对象后，在中间的文摘信息中可以用该对象的颜色高亮显示出来在文献中出现的位置，方便用户查阅。右侧展示该文献中计算得到的语义关系和句法关系，同样可以查看文中的具体位置。

示范系统所提供的语义化辅助阅读方式可以帮助用户直接查看最重要的知识点，直接定位重要知识所在的位置，引导读者优先阅读重点知识的段落和句子，从而提高对文献全文内容的阅读效率。

6 结语

本文基于 PubMed 的医学领域数据集合，采用较为成熟的数据挖掘、知识关系计算工具，研究提出了语义丰富化框架的设计模型，并通过构建示范系统进一步证明语义丰富化框架的优势和方法的可行性。总体来看，本研究主要从以下几个方面较大提升了语义化文献检索效果：

（1）语义识别技术将模糊的关键词匹配检索转变为具有语义特征的知识对象检索，提升检索精准度。辅助用户进行语义消歧，避免传统关键词检索出现的语义偏差。

（2）使用相关知识对象、事实关系、句子等更精准的语义知识代替全文检索，对科研人员判断该内容是否满足其检索需要有更大的帮助。通过语义关联定位真正知识所在的文献全文和段落。

（3）语义关联导航功能以数据统计方式揭示潜在关联知识，有助于发现隐含的知识关联信息，帮助科研人员发现新知识，探索学科交叉领域的新研究点，扩展科研人员的研究思路，辅助科技创新。

（4）语义化的辅助阅读高亮浮现重要知识点的位置，引导读者优先阅读重要段落和句子，提高文献全文内容的阅读效率。

在本文试验过程中，也发现了一些问题和不足之处，希望能够在未来的工作得以克服和改进：

（1）使用 clausIE 句法分析得到的 SPO 三元组关系是未经过规范的。本文已经使用领域词典对 S 和 P 进行二次规范，但是目前还没对谓词 P 进行规范。本文试验数据中未规范的谓词较为凌乱，对关联导航发现功能造成一定程度的影响，后续工作会构建谓词规范词表，或谓词语义识别方法对此改进和提升。

（2）计算得到知识对象之间的语义关系，和宽泛的上位词关联更为频繁，因此在揭示知识关系时出现宽泛的上位词较多，但大部分宽泛上位词对于专业领域的科研人员帮助不大，因此在未来工作中可以根据具体的科学领域，通过 TF/IDF 的方法计算知识对象的权值，过滤掉频繁出现又过于宽泛上位词，从而改进知识关联导航功能的效果。

参考文献：

[1] Pubmed [EB/OL]. [2015-10]. http://www.ncbi.nlm.nih.gov/pubmed

[2] Tan S, Zheng L. Methodology framework of knowledge organization system for scientific & technological literature. Library & Information. 2013;1:2-7.

[3] Semrep [EB/OL]. [2015-10]. https://semrep.nlm.nih.gov/

[4] Rindflesch, T.C. and Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics, 36(6):462-477.

[5] L. Del Corro and R. Gemulla. Clausie: Clause-based open information extraction. In Proceedings of the International World Wide Web Conference (WWW), 2013. 6

(作者 E-mail:xiej@mail.las.ac.cn)

作者贡献声明：

谢靖：语义丰富化检索模式设计及示范系统框架设计、研发，文章主要撰写人；

王敬东：数据语义标引，语义关系计算，文章语义标引章节撰写人；

吴振新：项目协调管理, 论文结构组织，论文版本修订；
张智雄：论文多维索引、语义检索的思路设计和指导；
王颖：示范系统数据组织和图形展示方案；
叶志飞：示范系统图形展示模块开发工作；