生物文献的本体建模及其在语义查询中的应用

李姣，朱小燕*

（清华大学 计算机系 智能技术与系统国家重点实验室，北京 100084）

摘 要： 由于现代生物技术的迅速崛起，大大加快了人们对基因组学，蛋白质组学等相关研究工作的步伐，使得与此相关的文献数量呈“爆炸式”增长。从海量生物文献中快速、准确地获取知的需求变得空前迫切。本文充分利用生物文献的结构化信息，对其进行本体建模。将 22,997 个生物学相关概念组织在一个上下位关系的层次结构中；从结构化的生物文献中抽象出研究人员、研究机构、基金等概念，并细致设计了上述概念的属性，及其之间的关系。我们将生物文献库 MEDLINE 中的文献作为实例导入到上述本体中，通过三个用例说明我们的本体在扩展生物领域词汇、查询关联复杂实体、跟踪生物领域发展状况等语义查询中的贡献。

关键词： 人工智能；知识表示；本体建模；语义查询

Ontology Modeling for Biological Literature and Application on Semantic Search

Jiao Li, Xiaoyan Zhu*

(State Key Lab of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing, China, 100084)

Abstract: The rapid development of biological techniques enhances the research of genomics and proteomics. Correspondingly, the amount of biological literatures is exploding with recording the new discoveries of gene, gene regulation, and protein-protein interaction etc. It is increasingly essential to acquire knowledge from the large-scale biological literature. We construct ontology for biological literature with its structured information considered. The ontology involves 22,997 biological related concepts organized in a hierarchical structure. Moreover, it abstracts important concepts from literature such as researcher, research group and grant etc. whose attributes and relationship are elaborately designed. One of the most popular biological literature resources, MEDLINE, is imported into our ontology as instances. Our ontology succeeds in three semantic search scenarios which yield that our ontology contributes in expanding biological term, searching object with complex relation and tracking biological domain development.

Keywords: artificial intelligence; knowledge representation; ontology modeling; semantic search

1 引言

随着高通量生物技术的发展，生物学的实验手段和研究方法均发生了巨大的变革，生物领域的数据呈指数速度增长。生物学文献作为成果展示和学术交流的主要方式之一，其数目之大，增长速度之快远远超过了其他学科领域。例如，隶属于美国国家生物技术信息中心的文献摘要库 MEDLINE，是世界上最大的、最具权威性的著名生物医学文献数据库[2]，2005 年的统计结果表明：它目前收集了全世界 4800 多种生物学及医学杂志上 15,000,000

基金资助：国家自然科学基金资助项目（60572084，60321002）。中国博士后科学基金资助项目（2005038088），感谢德国 Fraunhofer IPSI 的 Dr.Thomas Kamps 对本文部分工作的建议。

作者简介：李姣(1981-)，女，辽宁省本溪市，博士研究生，E-mail：jiao-li04@mails.tsinghua.edu.cn

*通讯作者：朱小燕，教授，E-mail：zxy-dcs@tsinghua.edu.cn

余篇文献信息，并且正以每个月超过万篇的速度增长。面对如此大规模的、快速增长的科学文献数据，如何从中检索到高质量相关信息的检索需求变得空前迫切，成为生物信息领域一个极具挑战性的课题。

借助于传统文本信息检索技术过去几十年的研究成果，生物文献检索取得了巨大的进展[3]。在以往的研究中，研究重点都集中在基于生物文献内容的信息检索包括：检索模型的建立，基于生物词典的查询扩展等。布尔模型和向量空间模型已经成功的应用到对外提供服务的生物学文本信息检索系统中（例如：PubMed[4]，E-BioSci[5]，Textpresso[6]等）；概率模型在生物学文本信息检索研究领域的国际评测中，体现出其在检索性能方面的优势并日益走向成熟；而语言模型的信息检索研究还处于起步阶段，目前仍然有很多问题需要进一步研究和解决[7]。生物术语学（Biological Terminology）研究者建立的一系列受控词汇表（Controlled Vocabulary，例如：MeSH[8]），本体（Ontology，例如：Gene Ontology[9]）等，通过查询扩展方式将生物领域知识引入检索任务中。

但是生物文本（尤其是生物文献，例如：MEDLINE 的数据）往往是有结构信息的， XML 格式 MEDLINE 数据中有对期刊名字，作者名字等信息的标注。如何将文本的内容与结构有机结合，提供高级的文本检索服务是本文研究的重点。对于生物学这样一个特定的研究领域，我们引入可以对概念和概念之间关系详细描述的本体（Ontology），对领域内不断产生的生物词汇和生物文献进行统一管理。对美国国家医学图书馆 NLM（National Library of Medicine）的医学主题词汇 MeSH（Medical Subject Headings）进行本体建模，并将其引入到生物文献（MEDLINE）的本体模型中。从基于生物本体的查询扩展、基于本体关系的语义查询和跟踪领域发展状况三个角度，说明了利用本体组织生物文献数据对语义查询的贡献。

下文各部分的安排如下：第二节详细描述了医学主题词汇 MeSH 和生物文献 MEDLINE 本体建模的过程。第三节通过三个用例说明基于上述本体的语义查询。我们在最后一节我们对本文的方法和后续的工作加以讨论。

2 本体建模（Ontology Modeling）

2.1 模型概述

本体（Ontology）一词源于哲学，在信息科学的研究中，本体用于将一个特定领域内的概念严格、无遗漏地表示在一个框架下。一个本体是对概念和关系的描述（概念化的详细说明），而这些概念和关系可能是针对一个特定领域而存在的[11]。

借助于本体的概念和关系描述，我们将美国国家医学图书馆的两大生物文本资源：医学词典 MeSH 和生物文献数据库 MEDLINE，统一到一个本体表示的框架下。

2.2 MeSH 生物本体建模

MeSH 是一个包含 22,997 个描述性词汇（descriptor）的医学词典，该词典由 15 个类别组成，例如：类别 A 为解剖学（anatomic）词汇；类别 B 为生物体（organism）词汇；类别 C 为疾病（disease）词汇；类别 D 为药物和化学制品（drugs and chemicals）词汇等等。每个类别进一步地被分为多个子类，构成了一个从一般到特殊的层次结构，最深的层次可以达到 11 层[8]。此外，MeSH 仍有 83 个修饰性词汇（qualifier），它们常常与描述词汇结合使用，从而对生物学概念加以准确描述。

我们为生物学词汇资源 MeSH 建立的本体 MeSHon，涵盖了 MeSH 中所有的描述词汇和修饰性词汇。描述性词汇被组织在一个 11 层的树状结构中，这些词汇之间的关系定义为上下位关系（subsumption）；修饰性词汇与上面的描述性词汇一起表达一个生物主题的某个方面，这些词汇之间是互不相关的；而二者中间的关系定义为修饰关系。在描述性词汇的树状结构中，一个概念可以出现 15 个不同的类别分支中，但是不能同时出现在一个类别分支内部。

与此同时，我们将 MeSH 词典中一些词条信息引入到上述本体的概念属性定义中，包括：唯一标号（UI），名字（Name），创建时间（DateCreated），内涵（Scope_Note），同义词（Synonyms）等。

2.3 MEDLINE 生物文献本体建模

美国国家医学图书馆的文献数据库 MEDLINE 的记录中有一个字段为<MH>，其中记录了体现该生物文献主

题的 MeSH 词汇，这就为 MeSH 和 MEDLINE 建立了天然的联系。将上述为生物词典 MeSH 建立的本体 MeSHon 引入到对 MEDLINE 生物文献的本体建模中，对生物文献的论述主题进行统一表示。图 1 给出了概念层上为生物学文献 MEDLINE 建立的本体：

Researcher 是一个拥有 *FirstAuthor* 和 *OtherAuthor* 两个扩展的可扩展概念（Extensible Concept），这样的设计体现了不同的 *Researcher* 可能在不同的 *Article* 中扮演不同角色的思想，例如：两个 *Researcher* 合作写过两篇 *Article*，根据贡献的多少他们分别担任 *FirstAuthor*，我们的本体模型就可以很好的涵盖该实例；

三个 *Organization* 与 *Article* 之间建立了不同的关系：*ResearchGroup* 拥有 *Article* 的所属，*FundAgency* 给 *Article* 提供 *Grant*，*PublishingHouse* 出版 *Article*；

MeSHon 中层次结构的描述性概念结合修饰性概念用以表示 *Article* 的主题。

为了使上述的本体表示更为灵活，我们采用反向关系（Reverse Relation）类型，例如：*Grant* 支持 *Article*，*Article* 得到 *Grant* 的支持等。此外，互反的关系对拥有不同的属性，例如："*is first author of*" 是复合类型的关系（Multiple Relation），而它的反关系 "*has first author*" 是单一类型的关系（Single Relation）。也就是说，一个 *Researcher* 可以多篇 *Article* 的作者，而一篇 *Article* 只能有一个 *Researcher*。同时，我们细致设计了上述概念的属性，包括：属性名称、类别、是否被索引、是否多次出现等。例如：我们为 *Article* 设计了名称为 *PubDate* 的属性，该属性的类型为时间类型 *Date(dd.mm,yy)*，它被索引且只出现一次，因此，用户的查询中可以涉及检索一篇在某日期之前、之后、之间发表的 *Article*。

基于上述概念层的本体，一篇生物文献经过解析后，相应于上述概念的实例（Instance）就会自动导入，并继承概念层的属性和关系。

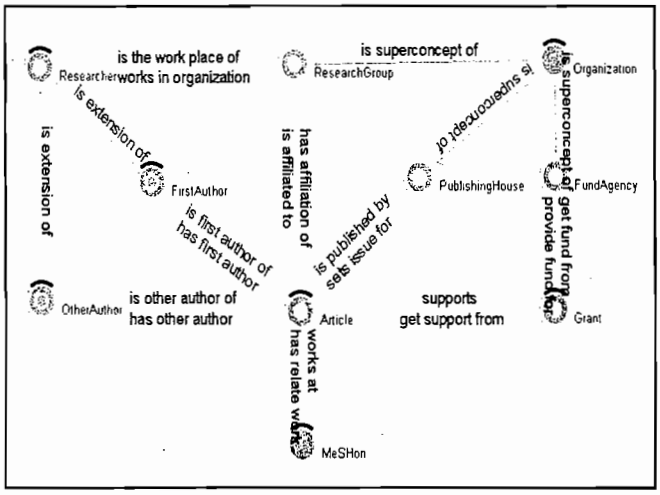
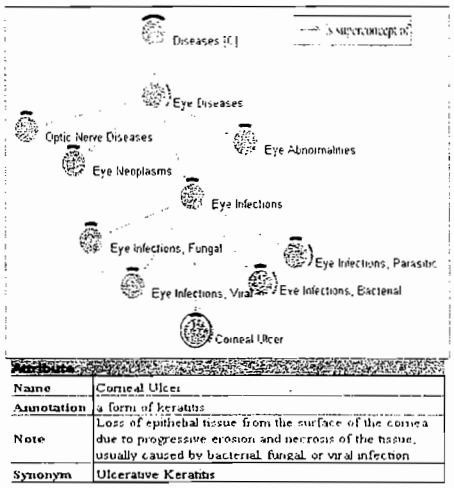


图1 生物学文献MEDLINE建立的本体在概念层的示例图



Attribute	
Name	Corneal Ulcer
Annotation	a form of keratitis
Note	Loss of epithelial tissue from the surface of the cornea due to progressive erosion and necrosis of the tissue, usually caused by bacterial, fungal, or viral infection
Synonym	Ulcerative Keratitis

图2 借助本体MeSHon进行查询扩展

3 语义查询（Semantic Search）

我们将 MEDLINE 的数据进行解析，作为实例导入到上面建立的本体中。分别通过眼病相关文献、人类基因组的基金支持、人类基因组计划发展状况查询用例，体现我们基于本体的语义查询在基于生物本体的查询扩展、基于本体关系的语义查询、跟踪生物领域发展状况的贡献。

3.1 基于生物本体的查询扩展

尽管 "Eye" 和 "Disease" 是两个简单的英文单词，但在生物领域内提出 "Eye Disease" 的查询需求，对其的正确理解将直接影响到检索性能。利用生物词汇本体 MeSHon 中概念属性和上下位关系，增强生物词汇的语义，从而改善检索性能。例如：一篇发表在 "Eye & Contact Lens" 期刊上，题为 "Ulcerative Keratitis in Contact Lens Wears" 的文章[12]，尽管全文中没有出现 "Eye Disease" 这个短语，但由于角膜溃疡（Ulcerative Keratitis）是眼

病的一种，人立刻就可以判断出该文章为眼病相关文章。我们的 MeSHon 通过以下方式使机器获得上述结论：如图 2 所示，概念“Corneal Ulcers”的同义词 Synomym 属性值为“Ulcerative Keratitis”；“Corneal Ulcers”是“Eye Infections”的子概念；“Eye Infections”是“Eye Disease”的子概念。由同义词的等价性和上下位关系的传递性，可以得到“Ulcerative Keratitis”是眼病的一种的结论。

这就为我们提供了一种查询扩展思路：利用生物本体中概念的属性和概念树状结构从一般到特殊的上下位关系，对用户的原始查询进行扩展。

3.2 基于本体关系的语义查询

在无结构化的文本中，检索出一个问题的答案是一个非常复杂的问题。通过对文本文件结构的解析，将其导入到特定本体中，建立起对象之间的关系。这就使得体现语义关系的查询成为可能。

图 3 给出了一个构造“检索出人类基因组计划的研究得到哪些机构的基金支持？”的查询。

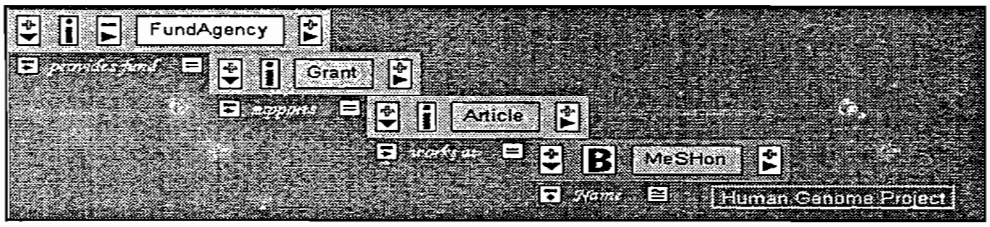


图3 查询人类基因组计划（HGP）研究受到基金支持

从第 2 节本体的建立中可见：我们的检索对象是提供基金的机构（*FundAgency*）；它提供基金（*Grant*）支持文章（*Article*）主题（*MeSHon*）相应领域的研究，其中指定 *MeSHon* 的 *Name* 属性值为人类基因组计划“*Human Genome Project*”。这样，系统就会将满足条件的 *FundAgency* 的实例包括：National Library of Medicine，National Institute of Neurological Disorders and Stroke 等提供基金的机构名称作为结果返回给用户。

3.3 跟踪生物领域发展状况

目前，生物领域是一个异常活跃的研究领域，按照时间序列跟踪领域的发展状况成为生物文本挖掘的一项研究内容。例如，通过对人类基因相关文献的统计，提出了人类基因的生命周期（Life Cycle）的概念。对生命周期中的鼎盛、衰退的研究总结出，某种基因（例如：CD4 与 p53）所收到的关注程度不单单由基因本身在细胞中所起到的作用决定的，而是同社会需求紧密相连的[13]。

我们对 1994-2003 年 10 年间有关人类基因组计划的文章发表、从事该项研究的机构、发表该项研究成果的期刊杂志、该项研究所获基金支持等情况做以统计（如图 4 所示）。

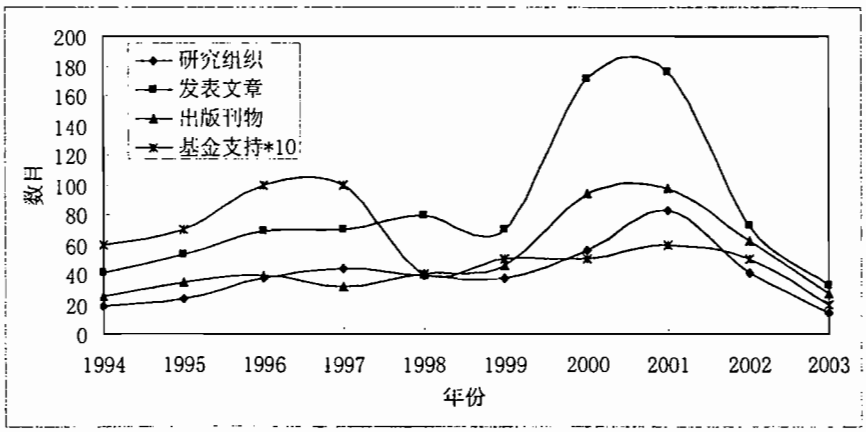


图4 跟踪人类基因组计划（HGP）在1994-2003年10年间的发展状况

可见，人类基因组计划 HGP 的发展在 2000 和 2001 年的时候到达了一个鼎盛的时期，这和人类基因组计划

的发展史：由于 Celera 公司的竞争，人类基因组计划较计划提前两年于 2000 年结束，直到 2001 年从事该项研究的科学家们开始大量发表他们工作所取研究成果，相吻合。而另一方面，基金支持的曲线体现出了研究“滞后性”这一特点，即，研究所取得的成果总是滞后于基金对其的支持。

通过上述的用例的分析总结出：基于我们构建的本体可以从多角度对生物发展现状进行跟踪，其结果验证了我们方法的有效性与客观性。

4 总结与展望

文本充分利用生物文献结构化的信息，借助于本体的概念和关系描述，将美国国家医学图书馆的两大生物文本资源：医学词典 MeSH 和生物文献数据库 MEDLINE，统一到一个本体表示的框架下。从而将基于结构的文本信息检索与基于内容的文本信息检索相结合，提供支持语义查询的高级检索服务。本文从生物领域文本角度对其进行了尝试性研究，通过 3 个用例说明基于生物文献本体建模的语义查询的可行性。

从另一个角度来看，文本尝试性研究的成功依赖于采用了高质量的数据，即：没有噪声的生物文献数据库 MEDLINE 的数据，以及借助了美国国家生物技术信息中心 NCBI 工作人员用统一的医学主题词汇 MeSH 对文献的手工标注。那么对于大量分布在各个生物刊物出版商的生物文献和各个生物研究所的技术报告，如何使用统一的词汇表对其进行标注；如何自动发现新的生物词汇，将其归入到生物医学词汇本体的适当的分支上；对语义查询返回结果的排序算法；以及对上述本体深层次的数据挖掘和知识发现等，都是值得我们进一步深入探讨的问题。

参考文献

[1] *GeneBank (access 2006).* http://www.ncbi.nih.gov/Genbank.

[2] *Fact Sheet of Medline (access 2006).* http://www.nlm.nih.gov/pubs/factsheets/medline.html

[3] Lars Juhl Jensen, Jasmin Saric, and P. Bork, *Literature mining for the biologist: from information retrieval to biological discovery.* Nature Reviews, 2006. **7**: p. 119-129

[4] *PubMed (access 2006).* http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed.

[5] Grivell, L., *E-BioSci: Semantic networks of biological information.* Information Services and Use 2003. **23**(2-3): p. 179-182.

[6] Muller HM, K.E., Sternberg PW., *Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature.* PLoS Biology, 2004. **2**(11): p. 1984-1998.

[7] William Hersh, et al. *TREC 2005 Genomics Track Overview.* in *Proceedings of 14th Text Retrireval Conference (TREC2005).* 2005. Gaithersburg, USA.

[8] *MeSH: Medical Subject Headings (acess 2006).* http://www.nlm.nih.gov/mesh/.

[9] Ashburner M, et al., *Gene Ontology: tool for the unification of biology.* Nature Genetics 2000. **25**(1): p. 25-29.

[10] Baeza-Yates, R. and G. Navarro, *Integrating contents and structure in text retrieval* SIGMOD Rec. , 1996. **25**(1): p. 67-79.

[11] Gruber, T., *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* International Journal Human-Computer Studies 1993. **43**: p. 907-928.

[12] Mela, E.K., Giannelou, Loanna P, John, Koliopoulos X, Sotirios, Gartaganis P, *Ulcerative Keratitis in Contact Lens Wearers.* Eye & Contact Lens: Science & Clinical Practice, 2003. **29**(4): p. 207-209.

[13] Hoffmann R, V.A., *Life cycles of successful genes.* Trends Genet, 2003. **19**(2): p. 79-81.