本体在生物医学领域中应用研究热点分析*

张 庆 吕少妮 轩 扬

（济宁医学院医学信息工程学院 日照 276826） （济宁医学院管理学院 日照 276826）

〔摘要〕 对 Medline 数据库收录的 8 年有关本体在生物医学领域应用论文中的高频主题词进行共现聚类分析，总结该领域研究热点，具体包括在生物信息学、临床医学、医学信息学以及人工智能中的应用共 4 个方面、8 个主题方向。

〔关键词〕 本体；生物医学；研究热点；聚类分析

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.01.014

Analysis of Study Hotspots of Ontology Application in the Biomedicine Field *ZHANG Qing, LV Shaoni, School of Medical Information Engineering, Jining Medical University, Rizhao 276826, China; XUAN Yang, School of Management, Jining Medical University, Rizhao 276826, China*

〔**Abstract**〕 The paper carries out co－occurrence cluster analysis of high－frequency subject terms in practical papers related to ontology in the biomedicine field that are collected by the Medline Database for eight years, summarizes study hotspots in the field, including application in bioinformatics, clinical medicine, medical informatics and artificial intelligence from four aspects and eight directions of subjects.

〔**Keywords**〕 ontology; biomedicine; study hotspots; clustering analysis

1 引言

本体最初是哲学领域概念，是对现实世界真实存在所做出的客观描述，20 世纪 90 年代本体概念被引入人工智能、图书情报和知识工程等领域[1]。本体是共享概念模型明确形式化规范说明[2]。由于生物医学领域庞大的概念和复杂的概念关系，应用本体表示知识概念进行知识组织显得尤为重要。本文对 Medline 数据库收录的有关本体在生物医学领域应用的论文中的高频主题词进行共现聚类分析，总结研究热点并对其进行分析。

2 数据来源与方法

2.1 数据来源

数据来源于 Medline 数据库。检索策略为（ontology［Title］OR ontologies［Title］）AND medline［sb］AND（" 2000/01/01" ［PDAT］: " 2017/12/31" ［PDAT］），共得到相关文献 1 810 篇。

〔修回日期〕 2018－09－18

〔作者简介〕 张庆，硕士，讲师，发表论文 8 篇，参编著作 3 部。

〔基金项目〕 济宁医学院科研计划项目“基于文本挖掘结果的本体的构建”（项目编号：0835/083501）。

2.2　方法

以 XML 格式套录检索结果，利用书目共现分析系统 BICOMB[3] 统计并抽取文献记录中的主要主题词与副主题词，按照出现频次由高到低进行排序，选取频次≥22 的 47 个主题词/副主题词作为高频词。其中出现频次最高的前 3 位主题词/副主题词分别为：受控词表、计算生物学/方法、软件。47 个高频词占所有与生物医学本体相关主题词的累计比例为 49.46%。对高频词在每篇文献记录中出现情况进行统计，形成高频词词篇矩阵。将词篇矩阵输入 gCLUTO 软件，采用系统聚类法对所得词篇矩阵进行聚类分析，结果可以反映出高频词之间的亲疏关系，根据高频主题词聚类结果以及主题词之间的语义关系总结出本体在生物医学领域中应用的研究热点。

3　结果

本体研究高频主题词共现聚类结果，见图 1。其中横轴代表文献，纵轴代表聚类的主题词/副主题词。两词聚集到一起的距离越短，关系越密切。

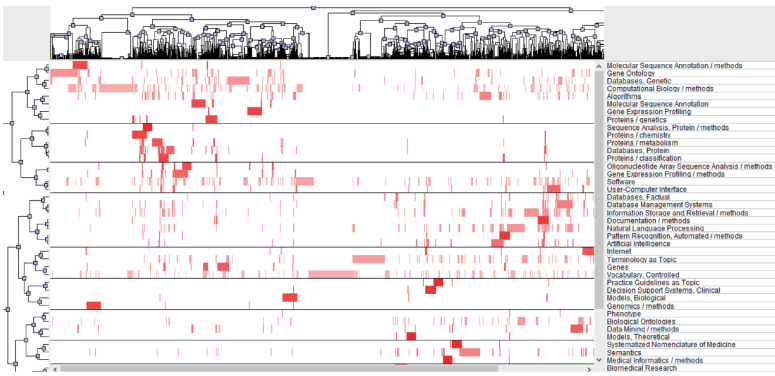
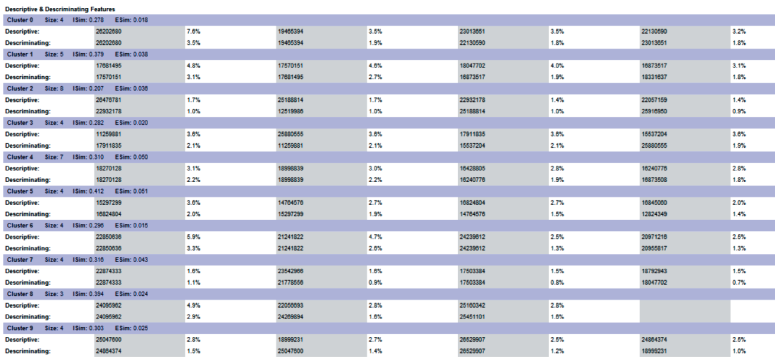


图 1　生物医学领域本体研究高频主题词共现聚类

首先，根据每类高频词的含义及其之间的语义关系总结出每类主题词所代表的研究热点，即生物医学领域本体的研究热点，如主题词蛋白质数据库（Protein Databases）和蛋白质/分类（Proteins/classification）距离较近，关系密切，先聚成一类；蛋白质/代谢（Proteins/metabolism）再与前面两个词合成一类，依此类推。通过分析这些主题词的语义关系能得出其所代表的类团含义标签，综合各个类别的类标签可以得出该主题的研究热点。其次，利用 gCLUTO 软件计算各类成员对聚类贡献率的指标（描述度和区分度），选择对每类形成贡献最大的来源文献作为表示该类内容的类标签文献[4]。通常选取描述度分值最高者作为该类的类标签文献，然后再对文献内容进一步分析，进而阐释该类研究方向的具体内容。类成员聚类贡献率指标，见图 2。



Descriptive & Descriminating Features								
Cluster 0　Size: 4　ISim: 0.276　ESim: 0.018								
Descriptive:	26202680	7.6%	19465394	3.6%	23013851	3.6%	22130590	3.2%
Descriminating:	26202680	5.5%	19465394	1.9%	22130590	1.8%	23013851	1.8%
Cluster 1　Size: 5　ISim: 0.379　ESim: 0.036								
Descriptive:	17681495	4.8%	17570151	4.6%	19047702	4.0%	16873517	3.1%
Descriminating:	17570151	3.1%	17681495	2.7%	16873517	1.9%	16331637	1.8%
Cluster 2　Size: 8　ISim: 0.207　ESim: 0.036								
Descriptive:	26470781	1.7%	25188814	1.7%	22932178	1.4%	22057159	1.4%
Descriminating:	22932178	1.0%	12519986	1.0%	25188814	1.0%	23918550	0.9%
Cluster 3　Size: 4　ISim: 0.282　ESim: 0.020								
Descriptive:	11259691	3.6%	25880555	3.6%	17911835	3.6%	15537204	3.6%
Descriminating:	17911835	2.1%	11259691	2.1%	15537204	2.1%	25880555	1.9%
Cluster 4　Size: 7　ISim: 0.310　ESim: 0.050								
Descriptive:	16270128	3.1%	16998839	3.0%	16428805	2.8%	16240776	2.8%
Descriminating:	16270128	2.2%	16998839	2.2%	16240776	1.9%	16873508	1.8%
Cluster 5　Size: 4　ISim: 0.412　ESim: 0.051								
Descriptive:	15297299	3.6%	14704576	2.7%	16824504	2.7%	16845080	2.0%
Descriminating:	16824504	2.0%	15297299	1.9%	14704576	1.5%	12824349	1.4%
Cluster 6　Size: 4　ISim: 0.296　ESim: 0.015								
Descriptive:	22850936	5.9%	21241822	4.7%	24239812	2.5%	20971218	2.5%
Descriminating:	22850936	3.3%	21241822	2.6%	24239812	1.3%	20955817	1.3%
Cluster 7　Size: 4　ISim: 0.316　ESim: 0.043								
Descriptive:	22874333	1.8%	23542966	1.8%	17503384	1.5%	16792943	1.5%
Descriminating:	22874333	1.1%	21778606	0.9%	17503384	0.8%	19047702	0.7%
Cluster 8　Size: 3　ISim: 0.394　ESim: 0.024								
Descriptive:	24866962	4.9%	22055093	2.8%	25160342	2.8%		
Descriminating:	24866962	2.9%	24266694	1.6%	25451101	1.6%		
Cluster 9　Size: 4　ISim: 0.303　ESim: 0.025								
Descriptive:	26047600	2.8%	19999231	2.7%	26529907	2.6%	24864374	2.6%
Descriminating:	24864374	1.5%	26047600	1.4%	26529907	1.2%	19999231	1.0%

图 2　类成员对聚类贡献率指标

4 分析

4.1 概述

通过对 Medline 收录的 8 年生物医学领域本体研究文献的高频主题词和副主题词进行共现聚类分析，可将该领域的研究热点分为 4 大方面、8 个主题。

4.2 本体在生物信息学中的应用

4.2.1 生物医学本体在蛋白质组学研究中的应用 包括主题词 Sequence Analysis, Protein/methods, Proteins/chemistry, Proteins/metabolism, Proteins/classification, Protein Databases。对蛋白质数据集进行功能注释分析对于解释高通量蛋白质组学的结果至关重要。生物医学本体正在成为蛋白质组学研究中的关键工具，用于蛋白质序列注释，预测蛋白质功能等。研究内容包括引入遗传相似性算法来寻找一组语义相似的基因本体术语，开发基于基因本体的蛋白质序列注释工具[5]；利用间接蛋白质相互作用来预测基因本体中蛋白质的功能[6]；利用基因本体和肽片段的几何聚类解析蛋白质中的功能重要区段[7]等。

4.2.2 生物医学本体在基因组学中的研究 该类涉及的主题词包括 Oligonucleotide Array Sequence Analysis/methods, Gene Expression Profiling/methods, Software, User - computer Interface。在生物医学本体中，基因本体（Gene Ontology, GO）已成为其中一种强有力的生物信息组织和加工工具。由于其被设计为物种中性，GO 非常适合跨物种使用，这意味着来自模式生物的功能注释可以转移到新测序物种中的推断的直向同源物中。即 GO 可以为具有未注释基因组的物种提供基因注释信息。研究内容包括与基因本体相关联的软件或应用程序的开发与应用，诸如 GO：TermFinder、JProGO、ChipInfo，用于提取基因注释和基因本体信息以进行微阵列分析[8-12]等。

4.2.3 本体在计算生物学中的研究 包括主题词 Molecular Sequence Annotation/methods, Gene Ontology, Genetic Databases, Computational Biology/methods, Algorithms, Molecular Sequence Annotation, Gene Expression Profiling, Proteins/genetics。利用本体来描述生物实体时可以通过对实体注释的含义相似性来评估两个实体之间的相关程度。语义相似性已成为验证生物医学研究结果的有用工具，如基因聚类、基因表达数据分析，分子相互作用的预测和验证以及疾病基因优先级。研究内容主要基于基因本体应用信息的语义相似度计算方法，获取不同基因产物生物特征的相似度。如基于本体语义相似性的功能分析工具 A - DaGO - Fun[13]；基于基因本体评估蛋白质功能相似度[14]；利用基因本体注释评估基因表达数据的聚类算法，用于解释基因表达数据以揭示共享共同功能属性的基因组[15]；基于基因本体注释的相似性预测蛋白质 - 蛋白质的相互作用[16]。

4.3 本体在临床医学领域中的应用

4.3.1 本体在临床决策支持系统中的应用 涉及的主题词包括 Practice Guidelines as Topic, Clinical Decision Support Systems, Biological Models, Genomics/methods。研究内容为开发本体用于临床实践指南及药物基因组学知识表示。如通过开发乳腺癌本体、基于指南要素模型以及患者本体在初级保健机构进行乳腺癌后续干预的临床决策支持系统[17]；开发网络本体语言（Web Ontology Language, OWL）和利用自动推理方法表示、分析和使用药物基因组学数据，使患者与临床上适当的药物基因组学指南和临床决策支持信息相匹配[18]；本体及其解决问题的方法在开发可共享临床指南中的应用，通过促进指南获取和执行，提高日常护理中可共享指南和决策支持系统的接受度[19]。

4.3.2 本体在分子生物学领域的应用 包括主题词 Phenotype, Biological Ontologies, Data Mining/methods, Theoretical Models。研究内容为通过生物医学本体查询和推断表型以用于临床基因诊断[20-21]。如根据患者表型对给定的一组基因进行排序。该算法通过在与每个基因相关的表型描述符和描述患者的表型描述符之间计算语义相似性来对基因进行排序。表型描述符术语取自人类表型本体（Human Phentypic Ontology, HPO），语义相似性源自每个术语的信息内容，可以相对于患者表型特征在基因列表内高度排列致病基因，以减少临床基因诊断的工作量[22]。又如结合开放生物医学本体（Open Biomedical Ontology, OBO）、自闭症本体与美国国立自闭症研究数据库（National Database for

Autison Research, NDAR), 采用描述逻辑和基于规则的推理方法, 从特定主题数据推断出高级表型抽象, 有助于研究人员进行数据分析[23]。

4.4 本体在医学信息学中的应用

4.4.1 与本体相关的术语词表研究 涉及的主题词包括 Internet, Terminology as Topic, Genes, Controlled Vocabulary。本体是受控程度最为严格、结构化程序最高的一种词表, 是知识表示的强大工具。领域本体描述的是特定领域中的概念与概念之间的关系, 提供专业学科领域中概念的词表以及概念间的关系, 能够合理有效地进行领域知识的表示。该类研究内容为构建词汇表供研究人员在研究过程中访问、浏览和利用。如开发健康术语/本体门户(HeTOP) 提供对健康术语和本体的轻松访问并可进行医学教学[24]; 通过编译 Gene Ontologies 生成描述分子生物学领域的结构化词汇表并将其应用于基因组表达分析中[25]。

4.4.2 本体在异构数据整合方面的应用 包含主题词 Systematized Nomenclature of Medicine、Semantics、Medical Informatics / methods。独立开发、结构各异的生物学数据库散落分布限制研究人员的具体研究。通过本体中的标准化术语不同数据集合的元数据可以被注释并进行术语统一, 进而消除异质性, 实现数据整合。该类研究内容主要体现在通过本体概念之间的简单术语匹配来解决整合异构知识源的问题[26-27]。如在 "Ontology Patterns - based Transformation of Clinical Information" 一文中作者提出一种灵活的转换方法, 使用语义内容模式来指导源数据和目标域本体之间的映射。作为用例, 该文展示如何使用 SemanticHealthNet 中提出的语义内容模式来转换有关药物管理的异构数据[28]。

4.5 本体在人工智能中的应用

包括主题词 Factual Databases, Database Management Systems, Information Storage and Retrieval/methods, Documentation/methods, Natural Language Processing, Pattern Recognition, Automated/methods, Artificial Intelligence。生物医学文献数量迅猛增长, 仅依靠人工检索阅读会消耗大量时间, 利用人工智能的方法能够有效地从生物医学数据库中提取相关知识进行研究进而提出新的实验假设, 得到新的科学结论。本体在人工智能中的应用主要体现在利用本体中的概念以及概念与概念之间的关系, 根据现用概念自动预测新概念, 结合算法自动实现多级注释、自动文本分类及聚类[29-31]。

5 结语

近年来本体广泛应用于生物医学研究中, 为学科领域中的类和关系提供标准标识符以及主题领域词汇表, 描述本体中类间关系含义的元数据, 机器可读的公理和定义, 便于计算机访问理解, 使其能够实现便于数据集成、数据访问和分析的应用程序。通过对高频主题词进行聚类分析, 可以总结出本体在生物医学领域中的应用主要集中在生物信息学、临床医学、医学信息学、人工智能 4 个方面。生物医学领域具有庞大的概念体系和复杂的概念关系, 使得本体对于该领域的重要性远远大于其他信息学领域。随着生物医学领域的发展, 本体将会被大量地用于知识和数据的表达与分析中, 向更高的覆盖范围、形式和整合方向发展。

参考文献

1 王向前, 张宝隆, 李慧宗. 本体研究综述 [J]. 情报杂志, 2016, 35 (6): 163 - 170.

2 Studer R, Benjamins V R, Fensel D. Knowledge Engineering: principles and methods [J]. Data & Knowledge Engineering, 1998, 25 (1 - 2): 161 - 197.

3 崔雷, 刘伟, 闫雷, 等. 文献数据库中书目信息共现挖掘系统的开发 [J]. 现代图书情报技术, 2008, 24 (8): 70 - 75.

4 史航, 高雯珺, 崔雷, 等. 生物医学文本挖掘研究热点分析 [J]. 中华医学图书情报杂志, 2016, 25 (2): 27 - 33.

5 Othman R M, Deris S, Illias R M. A Genetic Similarity Algorithm for Searching the Gene Ontology Terms and Annotating Anonymous Protein Sequences [J]. Journal of Biomedical Informatics, 2008, 41 (1): 65 - 81.

6 Chua H N, Sung W K, Wong L. Using Indirect Protein Interactions for the Prediction of Gene Ontology Functions [J]. Bmc Bioinformatics, 2007, 8 (Suppl 4): S8.

7 Manikandan K, Pal D, Ramakumar S, et al. Functionally Important Segments in Proteins Dissected Using Gene Ontology and Geometric Clustering of Peptide Fragments [J].

Genome Biology, 2008, 9 (3): 1 - 18.

8 Boyle E I. GO termfinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes [J]. Bioinformatics, 2004, 20 (18): 3710 - 3715.

9 Robinson PN, Wollstein A, Böhme U, et al. Ontologizing Gene - expression Microarray Data: characterizing clusters with Gene Ontology [J]. Bioinformatics, 2004, 20 (6): 979 - 981.

10 Brameier, Markus, Wiuf, et al. Co - clustering and Visualization of Gene Expression Data and Gene Ontology Terms for Saccharomyces Cerevisiae Using Self - organizing Maps [J]. Journal of Biomedical Informatics, 2007, 40 (2): 160 - 173.

11 Scheer M, Klawonn F, Münch R, et al. JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information [J]. Nucleic Acids Research, 2006, 34 (Web Server Issue): 510 - 515.

12 Zhong S, Li C, Wong W H. ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis [J]. Nucleic Acids Research, 2003, 31 (13): 3483 - 3486.

13 Mazandu G K, Chimusa E R, Mbiyavanga M, et al. A - DaGO - Fun: an adaptable Gene Ontology semantic similarity - based functional analysis tool [J]. Bioinformatics, 2015, 32 (3): 477.

14 Konopka B M, Golda T, Kotulska M. Evaluating the Significance of Protein Functional Similarity Based on Gene Ontology [J]. Journal of Computational Biology, 2014, 21 (11): 809 - 822.

15 Ma N, Zhang Z G. Evaluation of Clustering Algorithms for Gene Expression Data Using Gene Ontology Annotations [J]. Chinese Medical Journal, 2012, 125 (17): 3048 - 3052.

16 Maetschke S R, Simonsen M, Davis M J, et al. Gene Ontology - driven inference of protein - protein interactions using inducers [J]. Bioinformatics, 2012, 28 (1): 69 - 75.

17 Abidi S R, Abidi S S, Hussain S, et al. Ontology - based Modeling of Clinical Practice Guidelines: a clinical decision support system for breast cancer follow - up interventions at primary care settings [J]. Stud Health Technol Inform, 2007, 129 (2): 845 - 849.

18 Samwald M, Giménez J A M, Boyce R D, et al. Pharmacogenomic Knowledge Representation, Reasoning and Genome - based Clinical Decision Support Based on OWL 2 DL ontologies [J]. BMC Medical Informatics and Decision Making, 2015, 15 (1): 12.

19 Clercq P A D, Hasman A, Blom J A, et al. The Application of Ontologies and Problem - solving Methods for the Development of Shareable Guidelines [J]. Artificial Intelligence in Medicine, 2001, 22 (1): 1 - 22.

20 Shen Y, Zhang L. Gene Function Prediction with Knowledge from Gene Ontology [J]. International Journal of Data Mining & Bioinformatics, 2015, 13 (1): 50.

21 Wang H, Azuaje F, Zheng H. An Information Theoretic Approach to Assessing Gene - ontology - driven similarity and its application [J]. International Journal of Data Mining & Bioinformatics: 2014, 9 (2): 121.

22 Masino A J, Dechene E T, Dulik M C, et al. Clinical Phenotype - based Gene Prioritization: an initial study using semantic similarity and the human phenotype ontology [J]. BMC Bioinformatics, 2014, 15 (1): 248.

23 Tu S W, Tennakoon L, OConnor M, et al. Using an Integrated Ontology and Information Model for Querying and Reasoning about Phenotypes: the case of autism [J]. AMIA Annual Symposium Proceedings, 2008, (2008): 727 - 731.

24 Grosjean J, Merabti T, Griffon N, et al. Teaching Medicine with a Terminology/Ontology Portal [J]. Studies in Health Technology & Informatics, 2012, 180 (1): 949.

25 Blake J A, Harris M A. The Gene Ontology (GO) Project: structured vocabularies for molecular biology and their application to genome and expression analysis [J]. Current Protocols in Bioinformatics, 2002 (23): 1.

26 Lã3Pez - Garcã - A P, Lependu P, Musen M, et al. Cross - domain Targeted Ontology Subsets for Annotation: the case of SNOMED CORE and RxNorm [J]. Journal of Biomedical Informatics, 2014, 47 (2): 105 - 111.

27 Sǘnchez D, Solé - Ribalta A, Batet M, et al. Enabling Semantic Similarity Estimation across Multiple Ontologies: an evaluation in the biomedical domain [J]. Journal of Biomedical Informatics, 2012, 45 (1): 141.

28 Legaz - García M C, Martínez - Costa C, Miñarro - Giménez J A, et al. Ontology Patterns - based Transformation of Clinical Information [J]. Studies in Health Technology & Informatics, 2014 (205): 1018.

29 Fan J, Gao Y, Luo H. Integrating Concept Ontology and Multitask Learning to Achieve More Effective Classifier Training for Multilevel Image Annotation [J]. IEEE Trans Image Process, 2008, 17 (3): 407 - 426.

30 Lee J B, Kim J J, Park J C. Automatic Extension of Gene Ontology with Flexible Identification of Candidate Terms [J]. Bioinformatics, 2006, 22 (6): 665 - 670.

31 Yuan S T, Sun J. Ontology - based Structured Cosine Similarity in Document Summarization: with applications to mobile audio - based knowledge management [J]. IEEE Transactions on Systems, Man and Cybernetics, 2005, 35 (5): 1028 - 1040.