基于多属性规则的生物医学语义关系研究*

范少萍 安新颖
（中国医学科学院医学信息研究所，北京 100020）

摘要：生物医学领域文献指数级增长亟需快速识别出领域内核心且关键的语义关系，开展领域知识发现研究。本文借鉴DisGeNET基于规则的关系得分方法设计思想，在专家咨询基础上，提出生物医学领域语义关系具有可靠性、重要性和新颖性3种属性，设定每种属性对应的指标及定量计算方法。通过分析结直肠癌领域关系计算结果，验证本文所提关系得分方法在关系重要性排序方面的有效性。该方法考虑维度更加全面，为生物医学领域知识图谱构建、知识发现等提供有益参考。

关键词：语义关系；关系得分；关系排序；知识图谱；知识发现
中图分类号：G350 DOI：10.3772/j.issn.1673-2286.2021.01.003
引文格式：范少萍，安新颖. 基于多属性规则的生物医学语义关系研究[J]. 数字图书馆论坛，2021 (1)：18-23.

生命科学与医学是面向人民生命健康的科学，是关乎人类福祉的科学，科学探索与研究结论更强调科学性、可靠性、权威性。生物医学领域文献数量指数级增长，其中蕴含大量实体及相互间关系，对于构建领域知识图谱、厘清领域发展脉络、开展领域知识发现研究具有重要意义。

在实体间语义关系识别研究中，实体类型根据上下文语境具有唯一性，然而，实体间语义关系在不同语境与内容下有明显差异。如句子1“Aspirin use and survival after diagnosis of colorectal cancer”和句子2“Regular aspirin use after the diagnosis of colorectal cancer is associated with lower risk of colorectal cancer-specific and overall mortality, especially among individuals with tumors that overexpress COX-2”中，均提到了结直肠癌（colorectal cancer）与阿司匹林（aspirin），但二者关系在不同语境和内容下不同。句子1描述了阿司匹林与结直肠癌可能存在关系，在数据集[1]中关系标注为may be affected by；句子2明确说明结直肠癌诊断后定期服用阿司匹林与降低结直肠癌特异性和总体死亡率的风险有关，在数据集中关系标注为may be treated（decreased）by。此外，同一关系类型（如基因-疾病关系）甚至同一种关系在同一文献集合中重复出现，进一步强化了关系的确定性。因此，相关研究文献数量越多，关系数量越多，关系网络越复杂，亟需通过科学合理的方法识别核心关键关系，发现潜在有价值的关系，实现更细粒度的知识发现，从而为学科知识结构与知识关联构建、多源数据的深层次知识发现提供参考；同时也是辅助科研人员高效利用数据、发现新知识，以及提供智能精准的知识与情报服务的重要内容和发展方向[2]。

1 相关研究概述

现有语义关系排序方法主要分为两大类，即基于机器学习的方法和基于规则的方法。

1.1 基于机器学习的方法

李智恒等[3]、吴晓芳等[4]利用SemRep工具得到Medline数据库文献中不同语义类型对应的语义关系，采用KL散度、RlogF矩阵和PredScal函数相结合的方法计算语义关系得分，构建与疾病相关蛋白质以及蛋白

*本研究得到国家自然科学基金项目“面向精准医学的基因-疾病-药物语义关系抽取研究”（编号：71704188）资助。

质及药物等实体之间的联系。白洁[5]针对机器学习方法无法揭示语义层面深层次信息，构建关系本体开展语义关系抽取研究，并设计语义优先排序SPR算法，选取相关度最高的关系。孟祥福等[6]针对普通空间关键字查询通常会导致多查询结果的问题，根据空间对象之间的位置相近性和文本相似性，度量任意一对空间对象之间的位置-文本关系紧密度，提出基于概率密度的代表性空间对象选取算法。刘雷[7]基于复杂网络理论，选择多元关系排序问题开展研究，提出基于星型拓扑的异构概率超图模型，每个节点根据不同的关系类型，通过一定概率归属于不同的超边，为疾病预测可能相关联的基因。

然而，生物医学研究注重遵循证据，与循证医学（Evidence based Medicines）思想相同，即医生不是凭个人实践经验治疗患者，而是在个人丰富经验基础上，依据现有科学指导临床实践[8]。因此，机器学习算法在进行语义关系排序时仅对关系进行统计学计算，缺乏相关证据推演，对临床应用与指导价值的可靠性稍显不足。

1.2 基于规则的方法

基于规则的方法与基于机器学习的方法不同，并非通过大规模数理概率统计得到关系得分或排序，而是通过制定规则，基于规则约定或定量计算得到关系得分。生物医学领域应用此类方法的典型代表为DisGeNET数据库的关系得分计算方法。DisGeNET数据库通过整合已有多个基因/突变与疾病关联的数据库信息（如ClinVar、GWASDB等），利用机器学习方法从文献中获取相应的关联信息，同时，从关系来源数据库的数量和类型，以及支持关系的出版物数量方面，利用设计的方法计算关系得分（得分取值范围为0~1[9]），最终构建统一的基因/突变位点与疾病关联的数据库[10]。该数据库计算基因-疾病关系（Gene-Disease Associations，GDA）的方法见公式（1）。

$$S=C+M+I+L \quad (1)$$

其中，S代表关系最终得分。C代表关系是否被CGI、CLINGEN、GENOMICS ENGLAND、CTD、PSYGENET、ORPHANET、UNIPROT 7个数据库人工编审过，如果被3个以上数据库编审，则C=0.6；被2个数据库编审，C=0.5；被1个数据库编审，C=0.3；未被上述数据库编审，C=0。M代表是否包含RGD、MGD和CTD 3个数据库中的鼠类资源，如果包含，M=0.2，反之M=0。I代表是否包含HPO、CLINVAR、GWASCAT、GWASDB 4个人类表型、基因组等资源相关数据库中资源，如果包含，I=0.1，反之I=0。L代表LHGDN和BEFREE数据库中支持该关系的文章数量，如果有9篇以上文章支持该关系，L=0.1，如果支持文章数量不足9篇，则L得分为文章数量乘以0.01[9]。

可以看出，DisGeNET在计算GDA得分时重点依据该关系的来源数据库。来源数据库权威且数据库数据已得到编审与领域认可，则赋予较高得分。DisGeNET的这种GDA得分计算方法符合循证医学思想，利用相对充分的证据证明关系的科学性与可靠性，并利用关系出现频次证明其重要性。然而，上述关系得分计算方法仅考虑关系来源的权威性，忽略关系所在文献或文本所揭示出的其他有意义信息，如关系所在文本的出现时间可在一定程度上揭示关系的新旧程度，关系发表期刊水平可在一定程度上代表其对领域发展的贡献大小等。因此，须充分挖掘关系所在文本特征，界定关系多重属性，并遴选对关系排序具有关键作用的属性，从而设计更加科学、权威、全面的语义关系排序方法。

基于此，本文开展基于规则的语义关系研究，利用关系文本特征，梳理、总结关系属性，并设计科学可靠的属性计算方法，从而实现实体间语义关系的定量计算与排序，有利于识别重要语义关系，精简知识网络，发现潜在研究方向，对于推动生物医学研究高质量发展、辅助科研管理决策、实现研究成果转移转化等具有一定作用。

2 关系排序方法

为得到更科学合理的关系排序规则与定量计算方法，本节按照梳理关系属性、开展专家咨询、提出研究假设、设定计算指标、提出计算方法5个阶段分别开展研究。

（1）梳理关系属性。通过DisGeNET数据库关系得分计算维度可以看出，可靠性（来源数据库）与重要性（出现频次）是语义关系的两项重要属性。从科技文献老化的角度来说，旧有文献的使用频次将逐渐降低，让位于新生文献[11]，新生文献可能比旧有文献的研究内容更加新颖，对研究发展更具参考价值。标题与摘要揭示了文章的主要研究内容与最新发现，因此，发表在新

生文献标题或摘要部分的关系可能更新颖，关系根据其所在文献发表时间，具有新颖性。生物医学研究的最终目的是为临床服务，解决临床研究与决策中面临的各种科学问题，而临床试验是开展临床大规模应用的首要前提，因此，基于临床试验的应用性也应为关系的重要属性之一。基于此，生物医学领域语义关系应具有可靠性、重要性、新颖性和应用性4种属性。

（2）开展专家咨询。本文针对上述关系属性开展专家咨询，主要咨询科学计量学、计算生物学以及临床医学等领域相关专家和学者。专家和学者对可靠性、重要性与新颖性无太大异议；对于应用性，临床医学学者认为，文献中虽有部分经临床试验验证的研究结论与内容，但临床试验是非强制性要求注册并开展的项目，且研究文献中存在大量对临床有指导意义的基础研究，应用性指标的使用容易导致一些重要的基础研究中存在的语义关系被弱化，不建议纳入。因此，综合专家意见，本文最终确定关系属性为可靠性、重要性和新颖性3种。

（3）提出研究假设。结合上述分析结论，本文提出5种研究假设（本文所研究的关系均为正向支持关系）。

假设1：发表在高影响力期刊上的文献由于得到领域较高水平专家严苛评审，其所包含的关系可靠性相对较高。

假设2：比较毒理基因组学数据库（the Comparative Toxicogenomics Database，CTD）由于得到领域专家编审，数据更新及时，覆盖实体类型多样，在领域内应用广泛，被其收录并编审的关系具有较高可靠性。

假设3：标题是对文章关键核心内容的凝练，在标题位置出现的关系重要性更强。

假设4：关系三元组出现频次越高，其重要性越强。

假设5：文献出现时间越晚，越容易结合最先进思想与技术，在未来越可能有广阔研究空间，其所包含的关系新颖性越强。

（4）设定计算指标。基于本文提出的语义关系的可靠性、重要性和新颖性3种属性和研究假设，考虑数据获取的可靠性、便利性、可定量，设定每种属性所对应的指标及指标说明，如表1所示。与DisGeNET数据库相比，可靠性除考虑关系的来源数据库外，增加来源期刊影响因子，进一步强化关系的科学可靠。重要性在关系出现频次基础上，增加关系出现位置，进一步强化核心关键关系。此外，增加新颖性指标，从关系出现时间方面进行考量。

表1 语义关系主要属性、对应指标及指标说明

属 性	对应指标	指标说明
可靠性	权威性（Authority）	关系是否来源于CTD数据库
	影响力（Impact）	关系来源期刊的影响因子大小
重要性	核心力（Core）	关系在文献中所处位置（标题/摘要）
	频率（Frequency）	关系三元组出现频次多少
新颖性	新颖性（Novelty）	关系来源文献的发表时间

（5）提出计算方法。基于上述研究假设与指标设定，本文提出了关系排序计算方法，见公式（2）。虽然公式（1）和公式（2）有部分变量的字母表示相同，但含义大不相同，公式（2）利用每个指标英文首字母表示对应变量。

$$S=A+I+C+F+N \quad (2)$$

其中，S代表关系最终得分；A代表权威性，I代表影响力，二者合并为可靠性；C代表核心力，F代表频率，二者合并为重要性；N代表新颖性。权威性A根据关系三元组来源文献是否出现在CTD数据库计算，若出现在CTD数据库中，则A=0.2，否则A=0.1（CTD只是有针对性地对部分关系进行了人工审核，更强调关系的确定性，而未被纳入的关系，也已公开发表，得到领域专家认可，同样具有一定的意义，因此，不在CTD数据库也给予一定分值）。影响力I根据关系三元组所在期刊影响因子在所有同类关系中按从大到小排序的位置计算，最大值为0.2；若某类关系共来自10种不同影响因子的期刊，那么影响因子最高的期刊所描述的关系三元组影响力I=0.2，排序第10位的期刊所描述关系三元组影响力I=0.02；若期刊无影响因子，则影响力得分为0。核心力C根据关系三元组的出现位置计算，若出现在标题或同时出现在标题与摘要，则此关系三元组核心力C=0.2；若仅出现在摘要，则此关系三元组核心力得分为0.1。频率F根据关系三元组出现频次在所有同类关系中按从大到小排序的位置计算，最大值为0.2；若某类关系共有10种不同三元组，则出现频次最高的关系三元组的频率F=0.2，排序第10位的频率F=0.02。新颖性N根据关系三元组所在同类关系中文献的平均发表年份计算，若某类关系共有10种不同三元组，则平均发表年份最新的关系三元组新颖性N=0.2，排序第10位的新颖性N=0.02。如某关系三元组，其来源文献出现在CTD数据库，所在期刊影响因子排序为第1位，出现在文献的标题位置，关系三元组

出现频次排序为第1位，且发表时间排序第1位，则该关系的得分为S=0.2+0.2+0.2+0.2+0.2=1。本文将5个指标视为同等重要，在关系得分方法中所占比重相同，关系得分最高为1分。

3 实验与结果

3.1 关系来源与领域选择

本文计算关系来源于公开数据集[1]，该数据集主要包含肿瘤学相关研究的2 183条句子，标注了9种语义关系，分别为基因-疾病（GDA）、疾病-化合物（Disease-Chemical Associations，DCA）、化合物-基因（Chemical-Gene Associations，CGA）等，每类关系各3种。标注信息中含实体、实体关系、所在文献PMID等信息，可用于本文关系排序方法验证。

遴选标注文献中与结直肠癌相关文献。结直肠癌的发病率正逐年上升，在全球癌症中排名第3位。中国是全球结直肠癌每年新发病例数最多的国家。这种癌症是死亡率排名第二的疾病，正逐渐出现年轻化趋势。*CA-A Cancer Journal for Clinicians*发表的《2020年结直肠癌统计报告》（*Colorectal cancer statistics*，2020）[12]数据提示，年轻人肠癌发病率正在逐年增加，而老年人的发病率在降低，肠癌中位诊断年龄从2001—2002年的72岁降至2015—2016年的66岁。

3.2 计算结果与对比

根据公式（2）及各指标的计算方法，计算结直肠癌领域127条关系的得分情况，如表2所示。可以看出，得分最高的是5-Fluorouracil与结直肠癌可能存在may be treated（decreased）by的治疗关系，5-Fluorouracil是1962年上市的首个结直肠癌化疗药物，是结直肠癌化疗的首选药物之一[13]。在表2中，5-Fluorouracil与结直肠癌存在may be treated（decreased）by和may be affected by两种关系，且得分不同，主要是由于二者的核心力得分不同，也就是在文献中出现位置不同，如不纳入核心力指标，may be affected by的关系排序更靠前，但这一关系未能明确揭示5-Fluorouracil与结直肠癌是治疗关系还是致病关系。虽然两种关系同时出现，但经本文的关系排序方法计算后，may be treated（decreased）by的治疗关系排序靠前，且这一结果与文献分析结果一致，说明本文计算方法的有效性。

表2 关系计算结果（Top10）

序 号	实体1	实体1类型	实体2	实体2类型	关系名称	权威性得分	影响力得分	核心力得分	频率得分	新颖性得分	关系得分
1	colorectal cancer	Disease	5-Fluorouracil	Chemical/Drug	may be treated（decreased）by	0.200	0.158	0.200	0.200	0.178	0.936
2	TP53	Gene	colorectal cancer	Disease	is associated with	0.200	0.192	0.200	0.125	0.168	0.885
3	colorectal cancer	Disease	5-Fluorouracil	Chemical/Drug	may be affected by	0.200	0.164	0.100	0.200	0.181	0.845
4	APC	Gene	colorectal cancer	Disease	affects	0.200	0.163	0.100	0.150	0.183	0.796
5	colorectal cancer	Disease	fluorouracil	Chemical/Drug	may be affected by	0.200	0.164	0.200	0.080	0.143	0.787
6	TFPI2	Gene	colorectal cancer	Disease	is biomarker of	0.200	0.162	0.200	0.100	0.116	0.778
7	colorectal cancer	Disease	Aspirin	Chemical/Drug	may be treated（decreased）by	0.200	0.200	0.100	0.075	0.189	0.764
8	colorectal cancer	Disease	Aspirin	Chemical/Drug	may be affected by	0.200	0.196	0.200	0.040	0.124	0.760
9	colorectal cancer	Disease	resveratrol	Chemical/Drug	may be affected by	0.200	0.087	0.200	0.080	0.181	0.748
10	gastrin	Chemical/Drug	IEX-1	Gene	regulates	0.200	0.164	0.200	0.022	0.114	0.700

将本文方法得到的基因与疾病关系排序结果与DisGeNET数据库结直肠癌文献的基因与疾病关系得分结果进行对比，关系排序前5位的基因结果如表3所示。可以看出，与结直肠癌关系比较密切的基因（如APC、TP53）两种计算方法均排序靠前。其中，腺瘤性结肠息肉病（Adenomatous Polyposis Coli，APC）基因与结直肠癌关系类型为affects，通过文献阅读发现，该基因是结直肠癌抑癌基因，可在胚系和体系水平出现异常调节[14-15]，证明通过调节该基因可影响（affects）结直肠癌发生。此外，本文方法计算结果中K-RAS基因排序靠前，关系类型为is associated with。经查阅文献，RAS基因是第一个被鉴定出来的人类癌症基因，结

直肠癌的突变RAS基因主要是K-RAS。临床中较多研究聚焦K-RAS基因突变与结直肠癌转移和治疗的作用关系[16-17]，进一步证明该基因与结直肠癌多方面存在关联（is associated with）。但部分DisGeNET数据库排序靠前且与结直肠癌关系密切的基因（如CTNNB1），在本文计算结果中并未出现，可能与本文所选用关系计算来源数据集有关，数据集中包含的结直肠癌相关关系未覆盖所有结直肠癌基因。

表3 DisGeNET数据库和本文基因与疾病关系计算结果对比

计算方法与来源	排 序	基 因	关系得分
DisGeNET	1	APC	0.900
	2	CTNNB1	0.700
	3	MMP2	0.700
	4	TP53	0.700
	5	DPYD	0.600
本文	1	TP53	0.885
	2	APC	0.796
	3	TFPI2	0.778
	4	K-RAS	0.699
	5	IEX-1	0.642

化合物/药物与基因关系计算结果（Top5）如表4所示。本文识别出白藜芦醇（resveratrol）与K-RAS基因存在regulates关系。阅读文献发现，白藜芦醇是包括葡萄、花生和浆果在内的许多可食用品中的一种多酚化合物，其衍生物通过抑制结直肠癌致癌的K-RAS介导的信号通路来抑制HCT116细胞球体的生长，从而抑制具有K-RAS突变的结直肠癌细胞的增殖[18]。这一研究证明了白藜芦醇对K-RAS基因具有调控（regulates）作用，说明识别结果可靠。

4 结论与展望

本文遵循循证医学理念，借鉴DisGeNET基于规则的关系得分方法设计思想，在专家咨询基础上，提出生物医学领域语义关系具有可靠性、重要性和新颖性3种属性，基于研究假设，提出每种属性对应的指标及定量计算方法。通过结直肠癌领域关系计算与结果分析，验证了本文所提方法的有效性。与基于机器学习的关系排序方法相比，本文所提方法更注重相关证据推演，强调关系属性的多元性，更关注语义关系的临床应用与指导价值，更适用于生物医学领域的关系排序与语义关联构建。与DisGeNET数据库关系得分方法相比，本文所提方法考虑维度更加全面，纳入指标更加丰富，可以快速有效地识别关键关联，用于领域知识图谱构建；同时，由于加入新颖性属性，可用于发现新兴/潜在的语义关系，用于领域知识发现研究。因此，生物医学信息服务机构可利用本文所提关系排序方法，对已有数据库存储的语义关系进行计算，辅助科研人员识别并遴选实体间重要的语义关系，发现潜在有价值关系，更加清晰地厘清领域知识结构与内容演进，为相关研究提供选题参考与事实佐证。本文目前仅在一个数据集验证了方法的有效性，今后将在更多数据集进行验证，为生物医学领域重要语义关系识别与遴选、潜在有价值知识发现等提供有益参考。

表4 化合物/药物与基因关系计算结果（Top5）

序 号	实体1	实体2	关系名称	权威性得分	影响力得分	核心力得分	频率得分	新颖性得分	关系得分
1	gastrin	IEX-1	regulates	0.200	0.164	0.200	0.022	0.114	0.700
2	resveratrol	K-RAS	regulates	0.200	0.075	0.200	0.044	0.171	0.690
3	5-Aza-2'deoxycytidine	SPARC	regulates	0.200	0.139	0.200	0.022	0.114	0.675
4	fluoropyrimidine	K-RAS	has impact on	0.200	0.165	0.200	0.040	0.067	0.672
5	t-PTER	BCL-2	regulates	0.200	0.132	0.100	0.044	0.114	0.590

参考文献

[1] Corpus for relation classification in medical field [EB/OL]. [2020-11-10]. https://github.com/yangshuothtf/corpus_relation_classification.

[2] 胡正银，刘蕾蕾，代冰，等. 基于领域知识图谱的生命医学学科知识发现探析 [J]. 数据分析与知识发现，2020，4（11）：1-14.

[3] 李智恒，杨志豪，林鸿飞. 基于语义的疾病相关蛋白质知识抽取 [J]. 山东大学学报（理学版），2016，51（3）：104-110.

[4] 吴晓芳，杨志豪，林鸿飞，等. 基于语义关系的疾病知识提取系

统［J］. 计算机工程，2015，41（1）：284-288，295.

［5］ 白洁. 基于本体的实体关系抽取与检索［D］. 沈阳：东北大学，2012.

［6］ 孟祥福，张霄雁，赵路路，等. 基于位置-文本关系的空间对象top-k查询与排序方法［J］. 智能系统学报，2020，15（2）：235-242.

［7］ 刘雷. 基于异构超图的多元关系排序研究［D］. 大连：大连理工大学，2019.

［8］ 张鸣明，刘鸣. 循证医学的概念和起源［J］. 华西医学，1998（3）：6.

［9］ Original Data Sources［EB/OL］.［2020-11-08］. https://www.disgenet.org/dbinfo.

［10］ PIÑERO J，ÀLEX B，QUERALT-ROSINACH N，et al. DisGeNET：a comprehensive platform integrating information on human disease-associated genes and variants［J］. Nucleic Acids Research，2017，45（1）：833-839.

［11］ 庞景安. 科学计量研究方法论［M］. 北京：科学技术文献出版社，1999.

［12］ SIEGEL R L，MILLER K D，SAUER A G，et al. Colorectal cancer statistics，2020［J］. CA-A Cancer Journal for Clinicians，2020，70（3）：145-164.

［13］ SUSAN G，ARBUCK M D. Overview of clinical trials using 5-fluorouracil and leucovorin for the treatment of colorectal cancer［J］. Cancer，1989，63（S6）：1036-1044.

［14］ DOW L E，O'ROURKE KP，SIMON J，et al. Apc restoration promotes cellular differentiation and reestablishes crypt homeostasis in colorectal cancer［J］. Cell，2015，161（7）：1539-1552.

［15］ SA R，SONG H L，WEI M H，et al. The impact of APC polymorphisms on the transition from polyps to colorectal cancer（CRC）［J］. Gene，2020，740：144486.

［16］ 刘佳明，刘伟，徐达，等. RAS基因突变对结直肠癌肝转移患者肝切除术后预后的影响［J］. 中华肝胆外科杂志，2020，26（1）：1-5.

［17］ DIENSTMANN R，CONNOR K，BYRNE A T，et al. Precision therapy in RAS mutant colorectal cancer［J］. Gastroenterology，2020，158（4）：806-811.

［18］ OKAMOTO H，MATSUKAWA T，DOI S，et al. A novel resveratrol derivative selectively inhibits the proliferation of colorectal cancer cells with KRAS mutation［J］. Molecular & Cellular Biochemistry，2018：442（1/2）：39-45.

作者简介

范少萍，女，1986年生，博士，副研究员，研究方向：医学信息分析与科技评价。
安新颖，女，1978年生，博士，研究员，通信作者，研究方向：医学信息分析与科技评价，E-mail：an.xinying@imicams.ac.cn。

Study on Biomedical Semantic Relation Based on Multi-attribute Rules

FAN ShaoPing　AN XinYing
(Institute of Medical Information, CAMS & PUMC, Beijing 100020, China)

Abstract: With the rapid growth of biomedical literature, it is urgent to identify the key semantic relations in the field quickly, and carry out domain knowledge discovery research. Based on the design idea of DisGeNET and expert consultation, this paper proposes that biomedical semantic relation has three attributes: reliability, importance and novelty, and sets the index and quantitative calculation method for each attribute. Through the results and analysis of colorectal cancer relation, the effectiveness of the proposed relation score method in the ranking of relations and knowledge discovery is verified. The dimension of the proposed method is more comprehensive, which can provide a useful reference for biomedical knowledge graph and knowledge discovery.

Keywords: Semantic Relation; Relation Score; Relation Ranking; Knowledge Graph; Knowledge Discovery

（收稿日期：2020-12-05）