·医学信息研究·

整合词表概念优选名称自动生成机制研究*

李晓瑛 李丹亚 胡铁军 李军莲 钱 庆

(中国医学科学院医学信息研究所 北京 100020)

〔摘要〕 探讨多词表整合中概念优选名称自动生成机制，以UMLS超级叙词表为例进行详细阐述，并对该机制的正确性进行数据验证，分析评价其应用效果。

〔关键词〕 词表整合；概念；概念优选名称；一体化医学语言系统超级叙词表

The Research on Automatic Generation Mechanism for Preferred Name of Integration Vocabulary Concept *LI Xiao-ying, LI Dan-ya, HU Tie-jun, LI Jun-lian, QIAN Qing, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper discusses a generation mechanism for preferred name of concept in multi-vocabulary integration, and takes the University Medical Language System (UMLS) metathesaurus as an example to introduce this mechanism in detail, and proposes a testing method to verify and validate the correctness of this mechanism. Finally, it analyses and evaluats the application effect.

〔**Keywords**〕 Vocabulary integration; Concept; Preferred name of concept; University Medical Language System (UMLS) metathesaurus

1 引言

以概念为中心的多词表整合中，需要完成的一个重要工作是推荐概念的优选名称。因各个来源词表应用领域的差异，在表达同一概念时，往往存在多种不同表达方式的多个概念名称。目前，概念优选名称的生成方法主要为领域专家依据自身经验的主观推荐法、基于概念名称在生物医学文献中使用频率的选择法。这些方法的原理本质上相同，即从每个概念下多种不同表达方式中选择其一；在传统的纸本环境下，词表收录量相对较小，且概念名称的表达方式几乎固定，此类方法的确值得考虑。但在当今的网络环境下，研究人员可访问及获得的资源越来越多，词表的收录量也越来越大；并且随着研究人员在实际应用环境中对词汇的不断调整，概念名称的表达方式也经常发生变化，从而导致词表编制者需要重新选择概念优选名称，显然传统的概念优选名称生成方法已不能适应新时代的要求。本文以一体化医学语言系统（Unified Medical Language System，UMLS）[1]超级叙词表（Metathesaurus）为例，对概念名称标识及生成机制的作用原理进行详细阐述。

〔收稿日期〕 2012-11-20

〔作者简介〕 李晓瑛，博士，助理研究员，发表论文10篇。

〔基金项目〕 国家科技支撑计划“科技知识组织体系的协同工作系统和辅助工具开发”（项目编号：2011BAH10B02）。

2 概念优选名称自动生成机制研究

2.1 概念名称标识

UMLS 源自上百种医学词表，例如《医学系统化术语表 - 临床术语》（Systematized Nomenclature of Medicine - Clinical Terms, SNOMED CT）[2]、《医学主题词表》（Medical Subject Headings, MeSH）[3]，其超级叙词表便是通过广泛集成各部源词表中的生物医学概念、词汇及其涵义、等级范畴之后，所形成的多用途、多语种的词汇数据库。在超级叙词表中，来源于多部词表的表达同一概念的不同概念名称，通过概念惟一标识符（Concept Unique Identifier, CUI）联系起来，而全部概念名称及相应原始来源表信息存储于 MRCONSO 表[4]中，见表 1。其中，概念名称列于 STR（String）字段，SAB（Source Abbreviated）字段为相应的来源表缩写名称，TTY（Term Type）字段为来源表中术语类型的缩写（如 SNOMED CT 中‘FN’表示完全指定名称（Fully specified name），MeSH 中‘MH’代表主题词（Main Heading）），SUPPRESS（Suppression）字段为而概念名称的禁用标识，而前 8 个字段即为 UMLS 在整合各个来源表的概念名称时所生成的惟一标识符或标识，SAUI（Source Atom Unique Identifier）、SCUI（Source Concept Unique Identifier）及 SDUI（Source Descriptor Unique Identifier）为 UMLS 充分继承于来源表对概念名称赋予的标识符。

表 1 概念名称和来源（MRCONSO）

字段名	UMLS 注释	描述
CUI	Unique identifier for concept	概念惟一标识符
LAT	Language of term	语言种类
TS	Term status	词项状态
LUI	Unique identifier for term	词项惟一标识符
STT	String type	词串类型
SUI	Unique identifier for string	词串惟一标识符
ISPREF	Atom status - preferred (Y) or not (N) for this string within this concept	原词状态 - 概念的优选词串（Y）或非优选词串（N）
AUI	Unique identifier for atom - variable length field, 8 or 9 characters	原词惟一标识符 - 可变长度字段，8 或 9 个字符
SAUI	Source asserted atom identifier [optional]	来源表赋予的原词标识符（可选）
SCUI	Source asserted concept identifier [optional]	来源表赋予的概念标识符（可选）
SDUI	Source asserted descriptor identifier [optional]	来源表赋予的叙词标识符（可选）
SAB	Abbreviated source name (SAB)	来源表缩写名称
TTY	Abbreviation for term type in source vocabulary	来源表中术语类型的缩写
CODE	Most useful source asserted identifier, or a Metathesaurus - generated source entry identifier	来源表描述符，或超级叙词表所生成的来源表人口标识符
STR	String	词串（概念名称）
SRL	Source restriction level	来源表限制等级
SUPPRESS	Suppressible flag	禁用标识
CVF	Content View Flag	目录视图标识

如表 1 所示，UMLS 超级叙词表中的概念名称有四种惟一标识符：概念惟一标识符（CUI）、词项惟一标识符（Lexicon Unique Identifier, LUI）、词串惟一标识符（String Unique Identifier, SUI）及原词惟一标识符（Atom Unique Identifier, AUI），对应于三种标识：词项标识（Term Status, TS）、词串标识（String Type, STT）及原词标识（Is Preferred, ISPREF），这三种标识的赋值列于表 2 中。

表2 概念名称标识

标识	取值	UMLS 注释	描述
TS	P	Preferred LUI of the CUI	概念的优选词项
	S	Non - Preferred LUI of the CUI	概念的非优选词项
STT	PF	Preferred form of term	词项的优选形式
	VCW	Case and word - order variant of the preferred form	词项优选形式的大小写及字顺变体
	VC	Case variant of the preferred form	词项优选形式的大小写变体
	VW	Word - order variant of the preferred form	词项优选形式的字顺变体
	VO	Variant of the preferred form	词项优选形式的（其它）变体
ISPREF	Y	Preferred for this string within this concept	概念的优选词串
	N	Not preferred for this string within this concept	概念的非优选词串

表2 所列出的三种标识将提供重要的概念名称优选信息。首先，当某概念名称的词项标识 TS 赋值为‘P’，即可知该概念名称为同一概念下多组词项的优选词项；由于词项是经过忽略大小写、符号、字顺及屈折变形等处理后、具有相同原形一组概念名称的集合，具有相同原形的概念名称的词项标识符 LUI 相同，因此概念的优选词项往往有多个。其次，具有词串标识 STT 为‘PF’的概念名称，为词项的优选词串；而这组词项中的其他概念名称，即为优选词串的大小写、字顺等变体形式。再者，原词标识 ISPREF 的取值‘Y’或‘N’决定了概念名称是否为当前概念的优选词串，一个概念下包含多少种完全不同的词串（即 SUI），便有多少个优选词串。最后也是最重要的，这三种标识共同决定了概念的优选名称，即对于同一概念下具有不同表达形式的多个概念名称，只有词项标识 TS 赋值为‘P’、词串标识 STT 取值为‘PF’、原词标识 ISPREF 等于‘Y’的概念名称，才是概念的优选名称，而且是惟一的。

为了直观地从标识上理解概念优选名称的生成机制，本文引用了 UMLS 2011 AA 版中的一些概念名称，见表 3。因来源表及术语类型不同，在以‘C0000052’为代号的概念中，存在 8 个概念名称，且以 8 个不同的原词标识加以区别；通过三种标识可知，概念优选词项的词项惟一标识符为‘L0000052’，且其优选词串的词串惟一标识符等于‘S0007584’（对应于来自不同来源表的两个原词）；词项惟一标识符为‘L0006129’的一组词项，虽然并非概念的优选词项，但依然具有词串惟一标识符等于‘S0020479’的优选词串；原词标识为’Y’的概念名称共计有 5 个，代表了 5 种不同变体的概念优选词串；而概念的惟一优选名称却是原词唯一标识符为‘A0016536’的概念名称。

表3 UMLS 概念名称实例

CUI	TS	LUI	STT	SUI	ISPREF	AUI	SAB	TTY	STR	SUPPRESS
C0000052	P	L0000052	PF	S0007584	Y	A0016536	MTH	PN	1, 4 - alpha - Glucan Branching Enzyme	N
C0000052	P	L0000052	PF	S0007584	N	A0016535	MSH	MH	1, 4 - alpha - Glucan Branching Enzyme	N
C0000052	P	L0000052	VC	S0575717	N	A4769254	SNOMEDCT	OP	1, 4 - alpha - Glucan branching enzyme	O
C0000052	P	L0000052	VO	S0007578	Y	A0016529	MSH	PM	1, 4 alpha Glucan Branching Enzyme	N
C0000052	P	L0000052	VW	S2069199	Y	A1945338	MSH	PM	Branching Enzyme, 1, 4 - alpha - Glucan	N
C0000052	S	L0006129	PF	S0020479	Y	A0032514	MSH	EP	Branching Enzyme	N
C0000052	S	L0006129	VC	S0604824	N	A4773631	SNOMEDCT	IS	Branching enzyme	O
C0000052	S	L0006129	VW	S0038167	Y	A0054980	MSH	PM	Enzyme, Branching	N

2.2 概念名称优选级

从上面的分析结果可知，概念优选名称完全可用标识加以区别。例如，在 UMLS 超级叙词表中，概念优选名称可从每个概念名称的词项标识、词串标识及原词标识来判断。而 UMLS 在整合多个来源表的不同概念名称时，为这三种标识的赋值依据是来源表、术语类型及禁用标识的优选级，也就是在基于概念标识的概念优选名称自动生成机制中，为各个来源表的多种术语类型及禁用标识分配不同的优选等级，是生成机制最根本的作用原理。

在确定不同来源表的各个术语类型及禁用标识的优选等级时，UMLS 主要考虑了三种因素：来源表的学科覆盖率、更新频率以及概念名称在正规临床或生物医学用语中的使用率，最终确立的来源表及术语类型的优选级，记录在 MRRANK 表[4] 中。图 1 为 UMLS 2011 AA 版中 MRRANK 表的部分记录；对同一概念下的具有不同表达形式的多个概念名称而言，可根据各自的来源表缩写名称（SAB）、术语类型（TTY）及禁用标识（SUPPRESS），查找到相应的优选等级（RANK）；而具有最高优选级的概念名称，即为概念的优选名称。

2.3 标识的赋值机制

依据优选级表（如 UMLS 超级叙词表中的 MRRANK），便可为各个概念名称的词项标识 TS、词串标识 STT 及原词标识 ISPREF 分别赋值。其基本原理为：

```
RANK|SAB|TTY|SUPPRESS
0569|MTH|PN|N|
0568|RXNORM|MIN|N|
0567|MSH|MH|N|
0566|MSH|TQ|N|
0565|MSH|PEP|N|
0564|MSH|PEN|N|
0563|MSH|EP|N|
0562|MSH|EN|N|
0561|MSH|XQ|N
0560|MSH|PXQ|N|
0559|MSH|NM|N|
```

图 1 优选级（MRRANK）实例

首先，为概念优选名称的标识赋值：对于具有最高优选等级 RANK 的概念名称，其标识分别为 TS = ‘P’，STT = ‘PF’，ISPREF = ‘Y’；其次，为同一概念中其他概念名称的标识赋值：（1）为 TS 赋值：如果该概念名称与概念优选名称具有相同的词项惟一标识符 LUI，设置 TS = ‘P’；否则，TS = ‘PF’。（2）为 STT 赋值：在相同的词项惟一标识符 LUI 中，RANK 值最高的概念名称为优选词项，设置 STT = ‘PF’；对于和优选词项存在大小写及字顺、大小写、字顺或其他变体的概念名称，相应地设置 STT = ‘VCW’、‘VC’、‘VW’或‘VO’。（3）为 ISPREF 赋值：在相同的词串唯一标识符 SUI 下，RANK 值最高的概念名称为优选词串，设置 ISPREF = ‘Y’；对其余概念名称，设置 ISPREF = ‘N’。

3 数据验证

UMLS 超级叙词表中每个概念只有一个优选名称，且概念优选名称可通过词项标识、词串标识及原词标识来识别。为了验证这种生成机制的正确性，本文收集了 2011 年美国国立医学图书馆（National Library of Medicine，NLM）发布的及之后更新的 AB 共两个版本的 UMLS 超级叙词表中所有的英文概念名称及来源（MRCONSO）。验证算法的基本原理为：超级叙词表的概念总数应等于以 TS = ‘P’、STT = ‘PF’、ISPREF = ‘Y’为标识的概念优选名称的总数，可用结合了 SQL 查询语言的公式表示为：

| count (distinct CUI) | = | count (AUI) where TS = ‘P’ and STT = ‘PF’ and ISPREF = ‘Y’ |

其中，公式的左端用于计算概念总数，右端为概念优选名称的总数。

测试结果表明，2011 AA 版的英文超级叙词表总收词量为 7258 031，共含有 2404 336 个概念，而概念优选名称总数也为 2404 336；在 AB 版本中，超级叙词表的英文概念名称总数为 7651 245，其中概念总数为 2611 377，该数字也与概念优选名称的总数完全一致。

4 结论

本文以 UMLS 超级叙词表 （下转第 66 页）

求的习惯，成为具有优秀的思想文化、道德素质和创新能力的人才。

4.3.4 加强图书馆馆员的继续教育，注重人才培养 为适应新形式下图书馆发展的需要，图书馆支持鼓励馆员继续深造，建立了一套完整有效的馆员进修培训制度，采取走出去与请进来相结合的方式，开展多种形式的业务培训活动。连续多年坚持每年派出馆内各部门人员外出考察、进修，有针对性地学习其他图书馆的先进经验，先后去了北京、上海、广州等地的多所院校图书馆参观学习。同时，通过专题讲座、定期学术活动、每周的业务学习，拓宽馆员的知识面，使馆员能及时了解图书馆的发展趋势和学术动态，掌握多种工作技能，提高理论水平和工作能力。

5 结语

通过图书馆文化建设实践，深刻认识到：第一，图书馆文化建设是图书馆服务深化发展的产物，是图书馆发展到一定阶段的必然要求，需要图书馆的管理层和员工之间加强沟通交流，让全体员工不只是被动接受而是主动参与，这是开展图书馆文化建设的基础。第二，图书馆文化建设不只是图书馆某个党政工团组织或某个部门的额外任务，应当是图书馆的一个发展规划和整体目标，需要图书馆各级领导率先垂范、以身示教，全体馆员积极参与，这样的文化建设才会有生命力。第三，图书馆文化建设是一项长期的复杂而艰巨的系统工程，应采取长期的、有计划、有组织的塑造和建设才能取得明显成效。

参考文献

1 杨宇涵. 加强高校图书馆文化建设 全面营造育人氛围[J]. 图书馆建设，2002，(2)：23-24，31.

2 张映芳. 大学生参与高校图书馆文化建设探讨[J]. 广东药学院学报，2007，(1)：111-113.

3 王天珍. 高校图书馆文化建设问题研究[J]. 河北科技图苑，2009，22(5)：35-37，51.

4 蔡奎兵，曹平. 加强图书馆文化建设，提高读者整体素质[J]. 西域图书馆论坛，2006，(4)：30-31.

5 卢红. 新时期加强图书馆文化建设的思考[J]. 图书馆工作与研究，2012，(6)：73-75.

6 肖岚，叶翎. 高校图书馆人性化服务的探讨与实践[J]. 中华医学图书情报杂志，2010，19(10)：22-24.

7 原增，王虹菲. 图书馆的人本管理与和谐发展[J]. 中华医学图书情报杂志，2011，(10)：19-20，49.

8 赵李坚. 图书馆人本管理与文化建设[J]. 高校图书馆工作，2007，27(3)：78-80.

(上接第41 页)

为例，对概念名称标识及生成机制的作用原理进行了详细阐述。UMLS 超级叙词表，优选级 MRRANK 不仅确立了概念的优选名称，也决定了概念的优选词项、词项的优选词串以及概念的优选词串；而优选级又取决于各个概念名称的来源表、术语类型及禁用标识。在网络环境下，一方面，概念名称的表达方式会随时因实际应用需求的调整而发生变化，但由于术语类型不变，因而生成机制将不受影响；另一方面，虽然词表的收录量将随着可获得资源的增加而变得异常庞大，但因术语类型相对固定，这种基于标识的自动生成机制的工作量并不会增大。

基于概念标识的概念优选名称自动生成机制中，最核心的工作仍是仔细分析每个来源表的学科覆盖范围、更新周期以及概念名称的使用频率等，而这也是任何叙词表的编制者及多来源的超级科技词表整合者应着重研究的内容。

参考文献

1 U. S. National Library of Medicine. UMLS Home [EB/OL]. [2012-10-15]. http://www.nlm.nih.gov/research/umls/.

2 U. S. National Library of Medicine. SNOMED CT Browser [EB/OL]. [2012-10-20]. http://www.nlm.nih.gov/research/umls/Snomed/snomed_browsers.html.

3 U. S. National Library of Medicine. MeSH Browser [EB/OL]. [2011-12-17]. http://www.nlm.nih.gov/mesh/MBrowser.html.

4 U. S. National Library of Medicine. UMLS Reference Manual [Internet] - Metathesaurus [EB/OL]. [2009-11-20]. http://www.ncbi.nlm.nih.gov/books/NBK9685/.