

叙词表等同概念映射研究^{*}

刘伟 吴雯娜 王星

(中国科学技术信息研究所, 北京 100038)

摘要: 等同概念映射是叙词表融合中的必要环节之一, 本文从等同概念映射的准确性和效率两个方面, 讨论其中存在的问题及相应的解决方法。在映射准确性方面, 分析造成映射错误和映射丢失的原因, 包括用项不同、词形变化差异、非同义关系、同形异义; 在映射效率方面, 分别针对候选映射生成效率和人工审核效率提出相应的改进措施, 并进行实验验证。

关键词: 叙词表融合; 等同概念; 概念映射; 语义分析; 知识组织

中图分类号: G25

DOI: 10.3772/j.issn.1673-2286.2018.08.004

知识组织系统旨在表达特定学科或领域中概念和实体间的语义结构, 叙词表是其中代表性的知识组织系统之一。叙词表以概念为单位, 具有规范、简明的词间关系, 在传统信息组织中, 极大提高了文献的信息组织与信息检索能力^[1]。数字化和网络的发展, 推动了叙词表的出现、传播和应用。随着叙词表的增多, 对叙词表的合并与融合成了重要的应用需求和研究方向。通过对相同或相近领域叙词表的融合, 可以形成概念和语义关系更加全面的知识组织体系。《中国分类主题词表》是将《中图分类法》(第三版)与《汉语主题词表》进行融合的结果^[2]; 医学领域普遍使用的UMLS是将上百部生物医学领域的叙词表, 利用规范的关系进行融合, 是生物医学概念所构成的一部全面的叙词表^[3]。

叙词表的融合或归并, 是指不同叙词表中概念和语义关系的融合, 不同叙词表覆盖的学科领域既有重叠又有差异, 其中概念体系在详略、结构等方面也不相同, 使得融合过程充满挑战。常春等^[4]讨论了叙词表词间关系(等同关系、相关关系、等级关系)的合并方法; 吴雯娜等^[5]对叙词表融合中语义关系冲突的问题进行分析, 并提出相应技术方案; 邓盼盼等^[6]基于词表内聚合关系与映射类型的关联提出概念映射关系的处理方案, 并讨论了产生映射干扰的若干原因; 薛春香等^[7]对基于词形、结构和语料的术语映射方法进行总结, 对术语

映射中的方向性、不一致性、传递性等相关问题进行了讨论。作为另外一种知识组织系统, 本体间的集成、对齐等操作也需要在领域概念间建立起映射, Choi等^[8]对其中的映射问题和相应的映射技术做出回顾和总结。虽然本体和叙词表存在语言规范、组织结构等方面的区别, 但许多思想方法可用于叙词表的融合。

由于叙词表以概念为单位, 多部叙词表进行融合时, 首要步骤是建立叙词表间概念的映射。概念的映射类型可以分为等同、组代、多重属分等^[6], 其中等同概念间的映射是最主要的类型。不同叙词表的规模不一, 有的叙词表概念量非常大, 如EI叙词表(1993年版)有上万个概念。在多部词表之间大规模建立等同概念的映射, 完全采用人工方式在时间和成本上不现实。因此, 在实际操作中, 主要依靠计算机按照一定规则自动生成候选的概念映射, 然后由领域专家进行人工审核。目前候选映射的生成规则是基于字符串的匹配, 这种方式在有效性和效率方面都存在诸多问题和隐患, 本文将以英文叙词表中的概念为例, 讨论分析目前叙词表间概念映射存在的问题, 同时提出相应的解决方案。

1 等同概念映射中的问题分析

叙词表主要包含用、代、属、分、族、参等类型的词

^{*}本研究得到国家“十二五”科技支撑计划项目“《汉语主题词表》(工程技术版)与英文超级科技词表的映射研究”(编号: 2011BAH10B07)资助。

义关系,在多数情况下,用项(叙词)和代项(非叙词)之间表现为同义关系,成为建立等同概念映射的主要依据。在实际的映射操作中,一般把一个概念看作一个同义词群,用项作为该词群的优选词,代项作为该词群的非优选词。这是由于在一般情况下,用项和代项属于同义关系,而且可通过一定方法识别剩余非同义关系的用代项^[9]。对不同叙词表间的概念进行比对时,如果两个概念对应的词群优选词完全相同,那么这两个概念就形成一个候选映射,即两个概念可能为等同关系。随后,由领域专家审核决定候选映射是否成立。也有文献^[10-11]提出较为复杂的语义相似度计算方法,虽然理论上是一种可选择的方法,但在实际上计算代价较高,且有效性还有待提高。

叙词表之间进行等同概念映射时,可能的问题主要有映射错误和映射丢失两类。映射错误指不同的概念被映射在一起,映射丢失指等同概念没有被映射在一起。以下将对可能导致这两类问题的各种情况进行讨论分析。

1.1 用项不同

叙词表在编制时,对一个概念所有的同义词,应采用最广泛接受的术语作为用项(优选词),其他作为代项。早期的叙词表主要是领域专家依据其背景知识,依据各类工具书、专业术语等进行人工选词。较新的叙词表主要是以计算机为主,通过计算共现强度和词频等,自动从网络文献库中的关键词、网络热词及网络百科资源中选取^[12]。因此,叙词表编制时如果采用不同的叙词选择方法,同一个概念可能出现不同的用项术语。举例来说,早期叙词表一般会采用概念的全称作为用项,对应的缩略词作为代项,这主要是考虑到缩略词可能对应多个概念,含义不明确;在较新的叙词表中,如果一个缩略词被广泛熟知,且在叙词表领域中含义明确,而全称很少使用,该缩略词会被选为用项,全称被选为代项。如“human immunodeficiency virus”(人类免疫缺陷病毒)和其缩略词“HIV”。

通过上述分析得出,只通过用项匹配进行等同概念间的映射,显然会出现等同概念映射丢失的情况。为解决该问题,实际进行概念映射时,放宽了映射的限制,只要两个概念的同义词群有匹配的术语(不论用项还是代项),就将其作为候选映射。这样虽然提高了映射的召回率,但同时也使得准确率受到较大影响,造成人工审核的代价过高。

1.2 词形变化差异

英文词汇具有较丰富的语言形态变化,如复数、分词、过去式等。不同叙词表通常采用词汇的不同形式,这使得仅利用字符串匹配的方式会出现映射丢失的情况。即使在人工审核阶段,如果审核人员不熟悉一些词汇的特殊变化形式,也会导致映射丢失情况的出现。最常见的是单复数差异,如“document”和“documents”“tooth”和“teeth”等。为避免因词形差异导致的映射丢失,需要在映射前对词形作规范化处理,将构成术语的单词的不同变化形式统一化为标准形式,这个过程叫做原形化。

目前,原形化主要有词干提取和词形还原两种方式。词干提取是将术语的词干或词根抽取出来,但抽取结果可能是不具有实际意义的词,或者词干不一定能够表达完整语义。词形还原是通过形态分析,把单词的任一变化形式还原成一般形式,通常还具有实际含义的词典里的有效词。无论采用哪种方式,都有相应的计算机辅助工具可以直接实现。吴思竹等^[13]对常用的几种工具从功能、实现原理、实验结果等方面做了对比和分析。

单词形式的术语处理较为简单,但也有一定比例的术语是短语或其他形式。如“abuse of children”“child abuse”“abuse (children)”都对应概念“虐待儿童”,这需要对原形化工具的结果作进一步处理。张冰等^[14]对这方面的问题作了较细致的分析,提出一个基于原形化判断等同概念的流程。原形化虽然能够解决因词形差异造成的映射丢失问题,但同时会引入新的映射错误。如“cell”和“celling”是不同的概念,但原形化处理后会将其错误地映射在一起。

1.3 非同义关系

叙词表常因多种原因(如编制规则、人为错误等),使得用项和代项、代项和代项之间并非都是同义关系,还可能是近义、相关甚至是反义关系,如“reservoir”代项“water reservoir”和“seepage”代项“anti-seepage”。这些非同义词的代项将会导致映射错误的发生,使得不属于等同关系的两个概念作为候选映射。在被融合词表的数量较多时,这种映射错误会进一步被放大,增加人工审核的难度。以图1为例进行说明,假设概念A、B、C、D分别是来自不叙词表的概念,用粗体表示用项,其他为代项。假设概念A中用项a和代项b为非同义

关系(如反义关系),其他概念中用项和代项均为同义关系,这样产生的4个候选映射A-B、B-C、A-C、A-D,其中A-B和A-C就会是错误的。

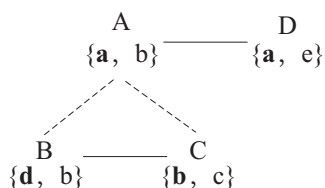


图1 非同义关系导致映射错误示例

1.4 同形异义

同形异义指形式相同而词义不相同的词汇,又叫做一词多义,如英语单词“set”就有上百个词义。在叙词表编制时,一般会选取单义词汇作为术语,尽量避免使用多义词汇,但仍然会有大量多义词汇不可避免地被编入叙词表。像“cell”有细胞、单元、电池等词义,在许多叙词表中都有收录,这类多义术语的存在显然会导致映射错误的发生,如概念“细胞”{cellula, cell, bioplast}和概念“电池”{battery, cell}。

无论是非同义关系还是同形异义造成的映射错误,都可以通过计算概念之间的词义相似度来判断是否为等同概念。概念在叙词表中以树状结构排列,分布在不同的“概念树”上,族首词为每棵“概念树”的根。高碧红^[10]和赖院根^[15]等利用叙词表中的等级关系,计算两棵“概念树”之间的相似度。当两部叙词表在学科领域和详略程度非常接近时,该方法较有效。在进行等同概念判断时,除通过叙词表自身固有的语义关系外,还可以通过概念的属性信息进行判断,如范畴分类、翻译、定义等;也可以利用叙词表中的概念翻译信息,对中文术语和英文翻译进行双向推导,实现对等同概念的识别^[16];从概念的定义中提取特征词,形成特征词向量,分析特征词向量之间的相关性,评估等同概念的可信度^[17]。如果候选映射中两个概念的范畴分类非常接近,那么正确的可能性非常大,否则可能性很小,这是因为一个术语在特定学科领域中的语义是唯一的。以上判断方法要根据具体情况而选用,因为叙词表的详略及包含的属性信息各不相同,这也是多部叙词表融合时,需要考虑的顺序因素。

2 等同概念映射中的效率问题分析

除有效性问题,效率问题也是等同概念识别的一个

重要问题,主要包括候选映射生成效率和人工审核效率。候选映射生成效率指如何快速生成所有的候选映射;人工审核效率指如何减少人工审核的次数。以下从这两个方面分别进行讨论分析,并提出可行的解决方法。

2.1 候选映射生成算法及效率分析

虽然候选映射是由计算机自动产生,如果不采用合适的生成算法,直接对不同叙词表中的任意两个概念进行比较,判断是否符合候选映射,将是一个非常耗时的工作。如对3个叙词表进行融合,各有1 000、5 000、10 000个概念,若采用直接比较的方式,概念比较次数为:1 000×5 000+1 000×10 000+5 000×10 000=65 000 000次。

理论上,如果对 n 个叙词表进行融合,各自的概念数量分别为 $m_1, m_2, \dots, m_n$,那么总的概念比较次数为 $\sum_{m_i \leq i, j \leq m_n} m_i \times m_j, i \neq j$ 。随着 n 的增大,概念比较次数将呈指数级增长。

因此,需要一些措施优化候选映射的生成过程,减少比较次数。在此本文提出一种优化的候选映射生成算法。①对进行融合的所有叙词表的术语进行原形化处理,将其结果称为术语原形。②对每个术语原形建立集合,集合中每个元素的结构为<原始形式,叙词表ID,概念ID,用代项>。③从每个术语原形对应的集合中生成候选映射,即集合中的任意两个元素对应的概念得到一个候选映射。如果集合中有 n 个元素,产生的候选映射的数量为 C_n^2 。④当前生成的候选映射中可能会存在重复,如两个概念中有多个术语原形重复,这一步可以与③同步完成,即每生成一个候选映射,就检查是否与已有候选映射重复。

在算法②中,原始形式是在叙词表中的原始形式,用代项指该术语在当前叙词表中是用项还是代项。具体实现时,可以与①同步完成:对一个术语原形化后,如果已有该术语原形,就其集合加入新元素;如果没有,就生成一个新的术语原形,集合中只有当前的新元素。利用该流程生成候选映射可以极大减少概念比较的次数,因为概念比较只在每个术语原形内发生。虽然生成术语原形需要额外的计算代价,但只是对每个叙词表的线性处理,与直接生成候选映射的比较次数相比,可以忽略不计。另外,候选等同概念不能包含所有的事实上的映射。如事实上等同的两个概念之间在原形化后不相同,就不会成为候选映射。

2.2 人工审核效率

在生成所有候选映射后,原则上需要审核人员对候选映射逐一审核,在审核过程中,可以将候选映射按照正确的可能性从高到低排序来提高审核效率。由于候选映射正确的可能性难以量化比较,这里指大概排序,而非绝对排序。下面是本文提出的一种排序方式:①候选映射中两个概念的用项精确匹配(忽略大小写差异);②候选映射中两个概念的用项原形化后精确匹配;③候选映射中两个概念的多个代项精确匹配;④候选映射中两个概念的多个代项原形化后精确匹配;⑤候选映射中两个概念的单个代项精确匹配;⑥候选映射中两个概念的单个代项原形化后精确匹配。

这样排序处理的优势是可以提高单个候选映射的审核速度,减少总体审核次数。一个候选映射准确的可能性越高,人工审核所花的时间就会越少,如排序①中的候选映射甚至可以节省人工审核的步骤。在部分情况下,前面的候选映射被审核后,后面的候选映射可以直接推出结果,这样可以减少审核次数。有3个候选映射A-B、B-C、A-C,如果经过人工审核A-B和B-C正确,那么A-C可以无须审核直接判断为正确;类似的,如果经过人工审核A-B正确,而B-C错误,那么A-C可以无须审核直接判断为错误。

3 实验分析

前文讨论分析了叙词表等同概念映射中存在的问题,并提出了提高效率的方法。为验证方法的有效性,在EI叙词表和INSPEC叙词表间进行验证。验证分为两个部分,一是用计算机自动生成候选概念映射的效率,二是人工审核候选概念映射的效率。

3.1 候选等同概念映射效率实验

EI叙词表有19 296个术语和9 926个概念,INSPEC叙词表有18 187个术语和10 075个概念。按照2.1小节中提出的候选映射快速生成算法,使用计算机共自动生成1 360个候选等同概念映射。第一步原形化需要对每个术语处理,共37 483 (19 296+18 187)次操作;第二步同样是对每个术语的处理,也是37 483次操作;第三步是得到的候选映射数量,即1 360次操作;第四步是重复候选映射,属于少量操作,对整体效率不会有影响。因

此,整个算法生成候选等同概念映射共需要76 326次操作。如果利用比较次数公式直接比较生成候选等同概念映射,则需要100 004 450次操作,显然远高于利用候选映射快速生成算法所需的操作次数。

3.2 人工审核效率实验

人工审核效率取决于候选映射中正确映射在所有候选映射排序中的位置,正确映射排序的位置越靠前,审核人员就会在候选映射中先审核到正确的映射,因此效率就会提高。本文采用经典的MAP(mean average precision)评价指标来比较在2.2小节提出的排序方式与随机排序方式。

MAP常用于评估搜索引擎的搜索结果排序,反映全部相关结果性能的单值指标。系统检索出的相关结果越靠前,MAP就应该越高。如果没有返回相关结果,则准确率默认为0。因此,能够用于对本文排序方法的评估。如10个候选映射中,正确的映射有4个,排名分别是1、3、4、8,那么对于MAP值的计算为 $(1 \div 1 + 2 \div 3 + 3 \div 4 + 4 \div 8) \div 4 = 3.92$ 。对1 360个候选映射经过人工判断,正确的映射共有1 249个,具体为:符合第一步的候选映射共976个,都是正确的;符合第二步的候选映射共141个,其中133个正确;符合第三步的候选映射共13个,其中6个正确;符合第四步的候选映射共16个,其中5个正确;符合第五步的候选映射共160个,其中101个正确;符合第六步的候选映射共有48个,其中29个正确。总体来看,正确的映射主要集中在排序靠前的部分,而错误的映射正好相反。按照每个正确的映射在排序中的位置,计算MAP的值为1 455.6。如果没有经过排序,将正确的映射随机排序(均匀分布),计算MAP的值为1 170.3。通过对比可以看出,本文的排序方法要优于随机排序,而且正确映射比值越小,优越性就会更加明显。

4 结语

随着知识组织系统数量的增多,映射与融合成为当前的热点研究趋势。等同概念映射是叙词表融合的重要环节,主要是通过计算机生成候选概念映射,由专家进行人工审核的方式完成。本文从有效性和效率两个方面对映射过程中的问题作出讨论,提出相应的解决方法,结果可以应用到多部叙词表间的概念映射,并在EI和INSPEC叙词表上进行了验证。

参考文献

- [1] 马张华. 信息组织 [M]. 北京: 清华大学出版社, 2009.
- [2] 国家图书馆《中国图书馆分类法》编辑委员会. 中图分类主题词表 (第二版) [M]. 北京: 北京图书馆出版社, 2005.
- [3] NIH [EB/OL]. [2018-01-01]. <http://www.nlm.nih.gov/research/umls/>.
- [4] 常春, 曾建勋, 吴雯娜, 等. 叙词表词间关系合并方法研究 [J]. 情报杂志, 2010, 29 (12): 117-120.
- [5] 吴雯娜, 王星. 叙词表融合方法研究 [J]. 中国图书馆学报, 2012 (4): 110-118.
- [6] 邓盼盼, 常春, 曾建勋, 等. 中英文词表概念映射关系及处理方案研究 [J]. 情报杂志, 2013 (10): 127-130, 192.
- [7] 薛春香, 乔晓东, 朱礼军. KOS互操作中的术语映射研究综述 [J]. 现代图书情报技术, 2010, 26 (2): 31-37.
- [8] CHOI N, SONG I, HAN H. A survey on ontology mapping [J]. SIGMOD Record: ACM SIGMOD (Management of Data), 2006, 35 (3): 34-41.
- [9] 王刘安, 常春. 用代传导中同义术语识别研究 [J]. 情报理论与实践, 2014, 37 (9): 97-100.
- [10] 高碧红, 吴雯娜. 叙词表中概念树相似度计算 [J]. 情报杂志, 2013, 32 (5): 155-157, 137.
- [11] 李荣, 杨冬, 刘磊, 等. 基于本体的概念相似度计算方法研究 [J]. 计算机研究与发展, 2011, 48 (z2): 690-695.
- [12] 浦墨, 郑彦宁, 赵筱媛, 等. 基于词共现关系强度和关键词词频的叙词选词方法探究 [J]. 图书情报工作, 2013, 57 (15): 121-125.
- [13] 吴思竹, 钱庆, 胡铁军, 等. 词干提取方法及工具的对比分析研究 [J]. 图书情报工作, 2012, 56 (15): 109-115, 142.
- [14] 张冰, 常春. 基于术语原形化的英文同义词群构建方法研究 [J]. 情报杂志, 2014 (7): 171-175.
- [15] 赖院根, 吴雯娜. 基于叙词表的概念语义相似度计算 [J]. 图书情报工作, 2009, 53 (8): 21-24, 33.
- [16] 宋培彦, 李静静, 赵星. 跨语言术语同义关系推荐方法及其实证 [J]. 现代图书情报技术, 2013 (5): 40-45.
- [17] 陈建超, 郑启伦, 李庆阳, 等. 基于特征词关联性的同义词集挖掘算法 [J]. 计算机应用研究, 2009, 26 (7): 2517-2519, 2532.

作者简介

刘伟, 男, 1976年生, 博士, 副研究员, 研究方向: 信息组织、文本挖掘, E-mail: liuw@istic.ac.cn。

吴雯娜, 女, 1972年生, 硕士, 副研究馆员, 研究方向: 知识组织。

王星, 男, 1977年生, 硕士, 工程师, 研究方向: 信息管理。

Research on Equivalence Concept Mapping among Thesauri

LIU Wei WU WenNa WANG Xing

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Equivalence concept mapping is a crucial step during the process of thesauri merging. The related problems and the corresponding solutions are discussed from effectiveness and efficiency: on the aspect of effectiveness, the reasons for mapping errors and mapping missing are analyzed, which includes the different usage items, term morphology, non-synonyms and homonymies; on the aspect of efficiency, the improvement measures are proposed for the generation of candidate concept mappings and manual review, which have been empirically verified.

Keywords: Thesaurus Merging; Equivalence Concept; Concept Mapping; Semantic Analysis; Knowledge Organization

(收稿日期: 2018-08-01)