<<知识组织

非相关文献知识发现初始集过滤方法的试验研究*

张云秋[1]　冷伏海[2]

[1]吉林大学公共卫生学院　长春 130021　[2]中国科学院文献情报中心　北京 100190

〔摘要〕在对现有非相关文献知识发现的初始集过滤方法进行分析的基础上,提出基于副主题词和基于共现语义群两种过滤方法。以 Swanon的早期发现之一为对照进行试验,考察经两种方法过滤后中间集 B的范围以及目标关联词和目标关联对的出现情况,以此作为评价其对 B影响的依据。结果表明两种过滤方法均可提高 B的质量,从而提高发现效率。

〔关键词〕非相关文献知识发现　初始集　副主题词　语义群

〔分类号〕G353.1

Experimental Research on the Filtering Methods of the Original Collection of Disjoint Literature-Based Discovery

Zhang Yunqiu[1]　Leng Fuhai[2]

[1] School of Public Health, Jilin University, Changchun 130021　[2]Library of Chinese Academy of Sciences, Beijing 100190

〔**Abstract**〕Based on the analysis of the existing filtering methods of original collection of disjoint literature-based discovery, this paper proposes subheadings-based filtering method and co-semantic groups-based filtering method. Then, an experiment is conducted comparing with one of Swanson's former discoveries. The size of B and the occurrence of the target terms and target relations are explored to evaluate their effect on B. The results of the experiment indicate that the two filtering methods can improve the quality of B and enhance efficiency of discovery accordingly.

〔**Keywords**〕disjoint literature-based discovery　original collection　subheading　semantic group

1986年,美国芝加哥大学 Swanson教授提出了基于非相关文献的知识发现[1]。ABC是其基本的发现模式,在应用中进一步演化为开放和封闭两个发现过程。开放过程是由 A开始,通过 B来寻找与 A具有潜在关联的 C。其中,A称为初始集,B称为中间集,C称为目标集。封闭过程是由假定 A和 C存在关联开始,来寻找连接 A和 C的可能关联 B。这里,A和 C均可称为初始集,B可称为中间集或目标集。由此可见,ABC三个集合的质量直接关系到发现的质量,其中涉及多方面的问题,本文仅对初始集过滤方法进行研究。

1　现有研究及其不足

对于非相关文献知识发现中初始集的形成,有研究者采用最简单的提问式来构建。研究者认为知识发现应是基于一个大的初始集,在发现过程的始端形成一个粗的范围宽泛的初始集是必要的,只有该集合的范围足够大,才不至于遗漏可能的潜在关联[2]。但是,非相关文献知识发现是以预期发现类型为前提的,具有方向性和目标性。一个粗的初始集会形成大量的中间概念,进而形成大量的目标概念,会大大增加筛选工作量,同时也会形成大量的虚假关联。因此,在形成初始集的过程中进行过滤是必要的。

目前采用的过滤方法包括一体化医学语言系统(Unified Medical Language System, UMLS)语义类型和停用词表。就 UMLS语义类型过滤方法来看,虽然语义类型在确定领域知识时很有意义,但是在实际应用中则需要从 134个语义类型中为每一个发现任务进行选择。其缺点是需要大量的人工参与,也会产生语义类型的歧异理解和错误映射。对于采用停用词表的过滤方法,其主要问题在于词表的长短。短停用词表只能滤掉常见词和无意义词,过滤效果有限;而长停用词表则可能产生过滤过度,滤掉一些有意义的词。而且针对不同的发现对象,停用词表的长短及具体的构成

*本文系教育部社会科学研究基金规划项目"非相关文献知识发现的理论、方法及应用的拓展"(项目编号:07JA870005)研究成果之一。

收稿日期:2008-10-03　修回日期:2008-12-15　本文起止页码:116-119,12　本文责任编辑:易　飞

词也会发生变化。因此,利用停用词进行过滤只能保证基本的过滤效果。

2 方法改进

2.1 基于副主题词的过滤法

非相关文献知识发现中初始集的形成,其实质是一个检索过程。因此,能够提高检索质量的方法,均应有利于构建初始集。虽然副主题词对于提高检索质量具有重要意义,但在非相关文献知识发现中尚未对其所起作用进行深入研究。

主题词是表达文献主题概念的规范性语词,主题概念间存在着各种逻辑关系。逻辑学上将概念上有关联但不如等同关系、等级关系严密的关系称为类缘关系或不确定关系,包括因果关系、影响被影响关系、应用关系和相关关系等[3]。由于这些逻辑关系划分的不确定性,常常导致虚假组配的产生和语法歧义。副主题词对主题词起方面限定作用,这种限定是对主题词自然属性的限定,通过这种限定使模糊的、不确定的逻辑关系明确下来,即副主题词能对主题概念间的类缘关系进行揭示。据此,对某一主题进行知识发现的过程中,针对发现的可能类型,分析与初始概念之间的逻辑关系,组配恰当的副主题词,从理论上讲能提高中间集和目标集的准确性,从而提高发现效率。

2.2 基于共现语义群的过滤法

UMLS主要由超级叙词表(metathesaurus)和语义网络(semantic network)构成[4]。其中,语义网络包括134种语义类型(semantic type, ST)和54种语义关系(semantic relation, rel)[5],用以标引超级叙词表中的概念及其相互之间的关系。因此,在非相关文献知识发现中常会利用语义类型进行过滤。但语义类型在应用中又显得过于复杂,为了减少其复杂性,有研究者提出了语义类型的上位概念——语义群(semantic groups, SG)[6],即按一定的原则将134个语义类型归为15个语义群。

语义群一经形成,它们之间必然会存在着语义关联。在UMLS的语义网络中,每一个语义关系rel是指两个语义类型ST_1和ST_2之间的关系。因为每一个语义类型都会归属于一个特定的语义群,因此,两个语义群之间的关系(SG_1, rel, SG_2)可以被认为是通过两个群中的语义类型之间的关系(ST_1, rel, ST_2)联系起来的。根据对UMLS语义网络的下载文件SRSTR的统计发现[7],15个语义群间的关系强弱不同。据此,在一个语义群对中,与两者共现的语义群也将存在着强弱不同的关系。

针对医学领域的特点,目前进行的非相关文献知识发现主要是疾病的可能致病因素或具有治疗作用的物质。因此,在分别计算语义群"疾病"(disorders, DISO)和"化学物质"(chemicals, CHEM)的关联语义群的基础上,计算其共现语义群及其关联强度,结果如图1所示:

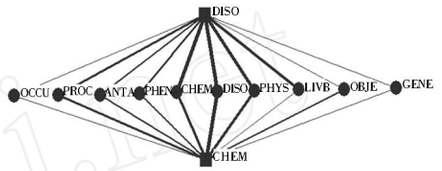

图1 与disorders和chemicals共现的语义群关联

图1显示,在DISO-CHEM语义群对中,有10个语义群与其共现,语义关系强弱不同,但这仅是从语义关系数量上来描述。非相关文献知识发现主要考察文献主题之间是否具有逻辑互补性和可推断性。因此,笔者进一步对上述DISO-CHEM语义群对中每一个与其共现语义群的具体语义关系进行分析,考察语义关系是否具有互补性和可推导性。结果表明,DISO、CHEM、PHEN(phenomena)、PHYS(physiology)、PROG(procedure)和ANTA(anatomy)6个强共现语义群,无论从语义关系数量还是从语义关系类型、具体的语义关系及其分布等方面,均存在着逻辑上互补和推导性。

因此,基于上述研究,本文提出疾病—物质类型的发现中,应采用上述6个语义类型进行联合过滤。

3 试 验

3.1 评价方法

试验以Swanson"偏头痛与镁"的发现为参照,理由是目前在Swanson的研究中,该发现对潜在关联关系阐述最为详细,并且该发现已经被后续实验所证实。本试验以该发现中得出的43个中间关联词和11个中间关联对作为标准[8],定义如下评价指标:

N_B = B列表包含的词和短语的总数

$Targ_T$ = B列表中出现的目标关联词的数量

$Targ_P$ = B列表中出现的目标关联对的数量

$Recall_T$(词查全率) = ($Targ_T$/43) ×100 %

$Recall_P$(关联对查全率) = ($Targ_P$/11) ×100 %

$Recall_{TP}$(综合查全率) = $Recall_T$ × $Recall_P$

$Precision_T$(词查准率) = ($Targ_T$/N_B) ×100 %

其中,N_B越大,B词总数越多,人工干预量就越大,知识发现的效率就越低。$Recall_T$、$Recall_P$和$Recall_{TP}$越高表明潜在关联遗漏的机率就越低,知识发现的效率越高。$Precision_T$越高表明在实际得到的B中具有发现价值的中间词越多,准确性越高,知识发现的效率越高。

3.2 试验1:基于副主题词的过滤效果

3.2.1 试验步骤 为了分析副主题词的过滤效果,分别制定如下检索式:

A_1 =“Migraine Disorders”[mesh] AND Migraine[ti] AND (“1960/01/01”[PDAT]:“1987/12/31”[PDAT])

A_2 = (“Migraine Disorders/blood”[MeSH] OR “Migraine Disorders/complications”[MeSH] OR “Migraine Disorders/drug therapy”[MeSH] OR “Migraine Disorders/physiopathology”[MeSH] OR “Migraine Disorders/therapy”[MeSH] OR “Migraine Disorders/urine”[MeSH] OR “Migraine Disorders/cerebrospinal fluid” OR “Migraine Disorders/metabolism”) AND migraine[ti] AND (“1960/01/01”[PDAT]:“1987/12/31”[PDAT])

C_1 =magnesium[MeSH] AND magnesium[TI] AND (“1960/01/01”[PDAT]:“1987/12/31”[PDAT])

C_2 =magnesium deficiency[MeSH] AND magnesium[TI] AND (“1960/01/01”[PDAT]:“1987/12/31”[PDAT])

其中,A_1和A_2是关于“偏头痛”(Migraine)的检索式。A_2较A_1的不同之处在于“偏头痛”主题词组配了“血液”(Blood)、“并发症”(Complications)、“药物治疗”(Drug therapy)、“病理生理学”(Physiopathology)、“治疗”(Therapy)、“尿”(Urine)、“脑脊髓液”(cerebrospinal fluid)和“代谢”(Metabolism)副主题词。选择这些副主题词,是考虑到偏头痛与镁之间的逻辑关联。其中,“血液”、“尿”、“脑脊髓液”是代谢的途径,是“代谢”副主题词的下位词,因此在这里均予以考虑。C_1和C_2是关于“镁”(Magnesium)的检索式。C_1未组配副主题词。在C_2中,并没有采用副主题词组配,而是选择了与副主题词作用相同的先组主题词——“镁缺乏”(Magnesium deficiency)。这是由于“deficiency”具有其特殊性,它既可以作为副主题词,也可以作为特定的主题词。“deficiency”作为副主题词可与任何表达物质的主题词组配,除了已被公认的缺乏状态和疾病(如“magnesium deficiency”)。

在与“偏头痛与镁”发现一致的数据库与时间段内,将上述检索式分别应用于开放和封闭的发现过程。在开放发现过程中,从A开始,因此分别选择A_1和A_2来获取B,比较其结果的不同。在封闭发现过程中,从A和C两端同时开始,其组合的检索模式包括A_1 + C_1、A_2 + C_1、A_1 + C_2和A_2 + C_2四种。

3.2.2 试验结果与分析 利用上述检索式分别构建初始集,经文本挖掘过程获得B,其各项指标如表1和图2所示:

表1 不同检索式对B影响情况一览表

检索式	N_B	$Targ_T$	$Targ_P$	$Recall_T$	$Precision_T$	$Recall_P$	$Recall_{TP}$
A_1	7190	43	11	100.00	0.60	100.00	100.00
A_2	4 080	43	11	100.00	1.05	100.00	100.00
A_1 + C_1	1 758	43	11	100.00	2.45	100.00	100.00
A_2 + C_1	1 351	43	11	100.00	3.18	100.00	100.00
A_1 + C_2	838	24	9	55.81	2.86	81.82	45.66
A_2 + C_2	718	23	9	53.49	3.20	81.82	43.76

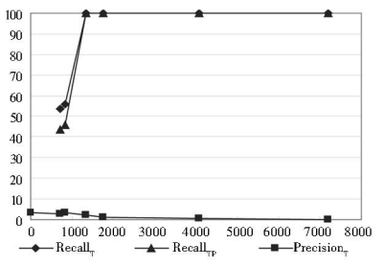

图2 不同检索式对B的影响效果

由表1和图2可见,以“偏头痛”单主题开始而获得的B包含7 190个词,若将其组配副主题词后,则B包含4 080个词,减少了43.25%,而43个目标词及11个关联对均未遗漏,词查准率提高0.45%。在以双主题(偏头痛与镁)开始而获得B的过程中,A、C均未组配副主题词时,B包含1 758个词;当A组配副主题词时,B减少23.15%,各项查全率均保持100%,词查准率提高0.73%。卡方检验(利用在两个B中目标词随机分布的概率模型)显示查准率呈显著提高。而C采用更专指的主题词后,B减少一半,但相应的目标词的查全率也降低近一半,关系对的查全率降低近19%,综合查全率降低超过50%。分析上述结果,A恰当组配副主题词能在不改变发现结果的同时,大幅度缩小B的范围。改由“magnesium deficiency”代替“magnesium”后,虽然B的范围进一步缩小,但却导致B的过度减少,从而影响发现效果。

3.3 试验2:基于共现语义群的过滤效果

3.3.1 试验步骤 在与“偏头痛与镁”发现一致的数据库与时间段内,分别采用开放和封闭的发现过程,对应用共现语义群进行过滤后产生B的情况进行统计分析。

3.3.2 试验结果与分析 开放发现过程中,对初始集经共现语义群过滤后,B的各项指标如表 2和图 3所示:

表 2 基于共现语义群过滤对 B影响情况一览表(开放)

语义群	N_B	$Targ_T$	$Targ_P$	$Recall_T$	$Precision_T$	$Recall_P$	$Recall_{TP}$
未过滤	7 190	43	11	100.00	0.60	100.00	100.00
DISO + CHEM + PHEN + PHYS + PROG + ANTA	4 370	43	11	100.00	0.98	100.00	100.00
DISO	1 059	17	6	39.54	1.61	54.55	21.57
CHEM	1 050	13	6	30.23	1.24	54.55	16.49
PROG	875	2	2	4.65	0.23	18.18	0.85
ANTA	601	2	2	4.65	0.33	18.18	0.85
PHYS	485	3	2	6.98	0.62	18.18	1.27
PHEN	300	6	4	13.95	2.00	36.36	5.07

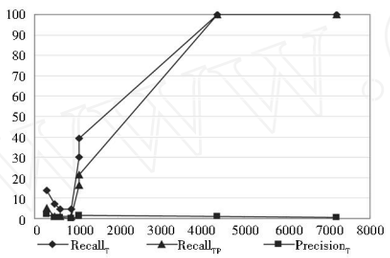

图 3 基于共现语义群过滤的 B的效果(开放)

封闭发现过程中,对初始集经共现语义群过滤后,B的各项指标如表 3和图 4所示:

表 3 基于共现语义群过滤对 B影响情况一览表(封闭)

语义群	N_B	$Targ_T$	$Targ_P$	$Recall_T$	$Precision_T$	$Recall_P$	$Recall_{TP}$
未过滤	1 738	43	11	100.00	2.45	100.00	100.00
DISO + CHEM + PHEN + PHYS + PROG + ANTA	1 045	43	11	100.00	4.11	100.00	100.00
CHEM	261	13	6	30.23	4.98	54.55	16.49
DISO	259	17	6	39.54	6.56	54.55	21.57
PROG	214	2	2	4.65	0.93	18.18	0.85
ANTA	146	2	2	4.65	1.37	18.18	0.85
PHYS	116	3	2	6.98	2.59	18.18	1.27
PHEN	49	6	4	13.95	12.24	36.36	5.07

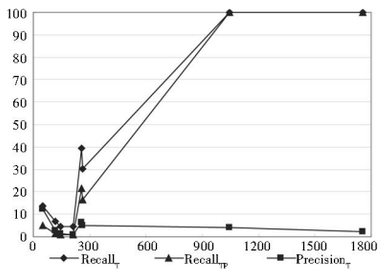

图 4 基于共现语义群过滤的 B的效果(封闭)

表 2和图 3显示,在开放发现过程中,未过滤产生 7 190个 B,包括 43个目标词和 11个关联对。采用 6个共现语义群进行联合过滤后,B减少了近 40%,但 43个目标词和 11个关联对并未遗漏,词查全率、关系对查全率和综合查全率均保持 100%,查准率提高 0.38%。表 3和图 4显示,在封闭发现过程中,未过滤产生 1 758个 B,包括 43个目标词和 11个关联对。采用 6个共现语义群进行联合过滤后,B减少超过 40%,但 43个目标词和 11个关联对并未遗漏,词查全率、关系对查全率和综合查全率均保持 100%,查准率提高了 1.66%。同时在两个发现过程中对 6个共现语义群分别过滤后的 B进行考察。结果显示,经每一个共现语义群过滤后目标关联词与关联对在 B中呈不均匀分布,其词查全率均在 40%以下,综合查全率仅维持在 20%以下。因此,单一的共现语义类型过滤意义不大。

综上分析,无论哪种发现过程,采用共现语义群对初始集进行联合过滤,均能在不改变发现结果的前提下,有效地缩减 B集合,提高 B的质量,进而提高最终的发现效率。

4 结 论

根据对上述试验结果的分析,证明恰当适度地组配副主题词,在不影响发现结果的同时,可有利于缩小初始集的范围,从而缩小中间关联集的范围,大大降低筛选的工作量,进而提高发现效率。但是,在副主题词使用过程中应该注意副主题词的恰当选取和适度选取。恰当选取主要是基于正确分析初始概念与可能的预期发现之间的逻辑关联,如试验中的偏头痛与可能致病或治疗物质之间主要是影响与被影响关系和因果关系,但若就此组配表达这些关系的副主题词则是不恰当的。因为,潜在致病或治疗物质对偏头痛的影响最有可能是通过作用于该疾病的病理生理过程的某一环节而产生作用或导致疾病的。从这一角度出发,组配相应的病理生理相关副主题词,则是恰当的。适当选取主要是副主题词的组配数量不可过多,对主题的限定不可过于严格。如果过于严格,则可能将那些处在边缘地带的主题删去。另外,由于词表本身及文献加工过程中的人为因素,均可能导致标引的不确切。因此,适当组配的原则是以删掉明显不符合要求的主题概念为准。

从基于共现语义群的过滤效果来看,经过强关联的共现语义群对初始集进行联合过滤后,可大大缩小中间集,即 B列表的长度,而发现结果没有改变,因此可大大提高发现效率。但是此过滤方法在实际使用中,共现语义群会随着发现类型的变化而不同。

(下转第 12页)

<<专　　题

以及分类使用知识经验有关系,这值得今后的实验分组给予关注。

4 结　语

通过上述对用户分类理解的理论与实验分析,我们获得了许多宝贵的第一手数据,这对我们今后的研究有着重要的借鉴与指导作用。

首先,我们尝试观测了用户对需求描述的心智模型,并尝试用实验进行测度与分析,获得了一些有意义结论:

用户在进行网络购物过程中,对需求商品会形成垂直概念,这对于使用相关分类体系是重要的。

用户在网络分类体系的选择和使用中,很大程度受其对商品描述的心理概念影响。

现有网站的网络分类体系设置与大多数用户的心智模型是存在有一定差距的。

其次,我们尝试观测了认知语境对用户分类理解的影响,并在一定程度上看到了网络界面符号体系中上下文的设置对用户符号理解是有作用的,这对我们如何设计用户易理解的界面符号体系是有参考意义的。

显然,我们的研究还有许多问题有待进一步探讨,尤其是有关实验的方法,比如用户心智模型揭示的观测与测度方法,网络界面语境、用户认知背景的科学观测方法,等等。

参考文献:

[1] 利奇. 语义学. 李瑞华,译. 上海:上海外语教育出版社, 1998.

[2] 胡壮麟. 当代符号学研究的若干问题. 福建外语 (季刊), 1999 (1): 1 - 9.

[3] 马张华,黄智生. 网络信息资源组织. 北京:北京大学出版社, 2007: 37 - 39.

[4] 张淑华. 认知科学基础. 北京:科学出版社, 2007.

[5] Sperber D, Wilson D. Relevance: communication and cognition. Oxford: Blackwell, 1996.

[6] 危辉. 语义问题与人工智能模型构造的系统观点. 心智与计算, 2008, 9(4): 14 - 15.

[7] 熊学亮. 认知语用学. 上海:上海外语教育出版社, 1999.

[8] Mey J L. Pragmatics: An Introduction. Foreign Language Teaching and Research Press Blackwell Publishers Ltd.

[9] Verschueren J. Understanding Pragmatics Foreign Language Teaching and Research Press, Edward Arnold (Publishers) Limited 2001.

[10] Cole C, Yang Lin, Leide J, et al. A classification of mental models of undergraduates seeking Information for a course essay in history and psychology: Preliminary investigations into aligning their mental models with online thesauri. Journal of the American Society for Information Science and Technology, 2007, 58(13): 2092 - 2104.

作者简介〕甘利人,女,1957年生,教授,博士生导师,发表论文 60余篇。

朱晶晶,女,1986年生,硕士研究生。

王静雯,女,1985年生,硕士研究生。

(上接第 119页)

参考文献:

[1] Swanson D R. Undiscovered public knowledge. Library Quarterly, 1986, 56(2): 103 - 118.

[2] Kostoff R N, Briggs MB, Solka JL, et al. Literature-related discovery (LRD): Methodology. Technological Forecasting & Social Change, 2008, 75(2): 186 – 202.

[3] 肖晓旦, 朱雷. MEDLINE副主题词特点及对医学主题词概念间逻辑关系表达. 情报科学, 2000, 18(6): 558 - 563.

[4] 方平. 试论一体化医学语言系统 (UMLS)超级叙词表的特点. 图书情报工作, 1998(10): 26 - 29, 41.

[5] 方平. 试论一体化医学语言系统语义网络的结构与特点. 情报学报, 1999, 18(2): 129 - 134.

[6] McCray A T, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Studies in Health technology and Informatics, 2001, 84(1): 216 - 220.

[7] SRSTR. [2008 - 06 - 20]. http://Semanticnetwork.nlm.nih.gov.

[8] Swanson D R. Migraine and magnesium: Eleven neglected connections Perspectives in Biology and Medicine, 1988, 31(4): 526 - 557.

作者简介〕张云秋,女,1972年生,副教授,博士,硕士生导师,发表论文 30篇,参编教材 10部。

冷伏海,男,1963年生,教授,情报研究部副主任,博士生导师,发表论文 60余篇,出版专著 (含)教材 12部。