中国图书馆学报(季刊)1992年第2期
ZHONGGUO TUSHUGUANXUEBAO

●河北机电学院图书馆主题法研讨组

中文科技文献主题法检索新探

一、句式法的提出

(一) 句式法概述。

随着科学技术的迅速发展，新生学科和边缘学科正在不断产生，计算机已被广泛应用。实现科技文献的计算机主题法检索是文献检索的总趋势。然而，由于现有主题法本身的局限性，事实上主题法还没有成为中文科技文献检索的主要方法。

我们认为影响主题法推广使用的主要因素有两个：一是系统性差；二是通过词表规范主题词，极大地加重了标引和检索工作的负担。解决这两个问题，就应在检索语言形式和标识单元方面进行新的探讨。

"汉语句式"是"句式法"的检索语言形式，是一种在汉语自然语言基础上加工后的人工语言。它兼容了专指性和一定的系统性特点。

"概念单元"是"句式法"的标识单元。它是每个特定主题中客观存在的概念单位，是一种取代了词表的规范主题词。

用"汉语句式"作为检索语言形式，以"概念单元"作为标识单元，按主题类集和检索文献的方法，就叫做"汉语句式主题法"，简称"句式法"。

(二) 检索语言设计的依据。

汉语具备着主题法检索语言的基本条件，并具有分类法一定的系统性特点。

(定语) 主语[状语]谓语<补语>(定语)宾语，这是标准的汉语单式句。从中可以分析出较理想的主题法检索语言的基本特征：

1.汉语没有词形变化。表示概念的词或词组不因条件影响而发生变化，使用方便。

2.汉语句中概念之间的关系，既是语法组配，又是逻辑组配，适于计算机准确检索。

3.标上语法职能符号的概念，可以独立，灵活运用，符合主题法检索语言的特点。

4.汉语句中各个不同的语法成份，形成了一个个自然的概念单位，即标识单元。

5.汉语句中的语义层次关系，可以通过以主语为首的主次地位层层展开，具有一定的系统性。

当然，汉语要成为一种检索语言，必须简明、准确，符合文献类集和检索的规律。我们根据主题法的基本要求，从文献的标引和检索实际出发，将汉语习惯语进行了改造和加工，在汉语有关句式的基础上，设计出了一种标定语法职能符号的检索语言形式——"汉语句式"：(限词) 主词[状词]<适词>面词述词宾词《缀词》。

句式中的限词、主词、状词、述词和宾词，其语法作用和标定符号，基本类似于汉语中的定语、主语、状语、谓语和宾语；适词相当于汉语中的介词结构状语；面词则包含汉语中的被动句，因果句等；缀词表示文献的类型、载体，类似分类法的部分复分号。

二、文献主题的标引方法

(一) 基本标引方法。

标识文献主题的汉语句式，分主词句和主述句两大类。

1.主词句，即揭示事物对象主体的句式。

(1) 主词，表示事物主体的概念。标定符号为"═══"。主词可以用名词或名词化的动词表示。如用镗床、热处理两个独词句式，分别标识主题为"各类镗床方面问题"、"各种热处理方面问题"的文献。这种句式揭示的主题，内涵浅，外延广，覆盖文献面最大。

(2) 限词，对主词修饰或限定的概念。标定符号为"( )"。如用(金属切削)机床、(同步)电动机两个句式，分别标识主题为"金属切削机床"、"同步电动机"的文献。这种句式，限词用的越多，事物主体越具体。

2.主述句，即表示事物主体动态、性态的句式。

(1) 述词，表示主词动态和性态的概念。标定符号为"____"。述词除用动词、形容词表示外，还

可以用动名词、性态名词表示。如用（螺纹加工）机床检测、模具热处理、汽车结构、可控硅性能几个句式，分别标识主题为“螺纹加工机床的各种检测”、“各类模具的各种热处理”、“各类汽车的结构”、“可控硅的各种性能”等文献。

(2) 状词，表示述词的手段、方式、使用的工具、范围等的概念。标定符号为“[ ]”。如用铸件[消除应力]退火、电视机[示波器]检测、铜[物理]性能几个句式，分别标识主题为“各种铸件的消除应力退火工艺、理论”、“电视机用示波器检测方法”、“铜的物理性能”等文献。

(3) 宾词，标定符号为“～～”。

a.表示主词通过述词影响或涉及的对象。如用电磁场干扰通讯句式，标识主题为“电磁场对通讯干扰的原理、预防”的文献。

b.表示主词演变的趋向或结果。如用电动机改装发电机句式，标识主题为“电动机改装成发电机的理论、方法”的文献。

(4) 适词，标定符号为“< >”。

a.表示主词所处的环境或具备的条件。如用钢<高温>塑性、铝<真空>[导电]性能两个句式，分别标识主题为“钢在高温中的塑性”、“铝在真空中的导电性能”的文献。

b.表示主词被应用的方面。如用超声波<医疗>应用句式，标识主题为“超声波在医疗方面应用”的文献。

c.表示主词发展、演变的来源。如用推土机<坦克车>改装句式，标识主题为“推土机由坦克车改装”的文献。

(5) 面词，标定符号为“…”。

a.表示主词被支配、影响的对象。如用钢氧腐蚀句式，标识主题为“钢被氧腐蚀的原理、防治”的文献。

b.表示现象产生的原因。如用铸件浇注裂缝句式，标识主题为“铸件因浇注问题而产生裂缝”的文献。

c.表示主词的局部或某方面。如用机床变速箱设计句式，标识主题为“各种机床变速箱方面的设计”的文献。

这种句式是为了将面词所表示的概念归入主词概念类下；遇到主述搭配不合理时，则不能使用。

最后一个是缀词，它表示文献的类型或载体，标定符号为“《 》”。缀词在主词句和主述句中都可能使用，如（高等）数学《题解》，英语会话《磁带》。

上述句式在表征主题含义方面，虽有重复现象，但却利于多途径检索。

(二) 同一主题的多途径标引。

为了满足读者多途径检索的需要，则将可能首先入手检索的概念分别充当主词，组成几种不同的句式来标识同一主题。

1.名词化的述词可做主词。如主题为“齿轮热处理”的有关文献，可同时用齿轮热处理、(齿轮)热处理两种句式标识。

2.表示工具、器物的状词可做主词。如主题为“用万用表检测电路”的有关文献，可同时用电路[万用表]检测、万用表检测电路两种句式标识。

3.表示主词局部或某方面的面词可做主词。如主题为“电机转子加工”的有关文献，可同时用电机转子加工，(电机) 转子加工两种句式标识。

4.表示被动关系的面词可做主词。如主题为“铁被硫酸腐蚀”的有关文献，可同时用铁硫酸腐蚀、硫酸腐蚀铁两种句式标识。

5.表示主词演变来源的适词可做主词。如主题为“发电机由电动机改装”的文献，可同时用发电机<电动机>改装、电动机改装发电机两种句式标识。

6.表示被应用方面的适词可做主词。如主题为“计算机在工业方面的应用”的有关文献，可同时用计算机<工业>应用、工业应用计算机两种句式标识。

7.表示原因的面词可做主词。如主题为“部件因疲劳而造成损伤”的有关文献，可同时用部件疲劳损伤、疲劳损伤部件两种句式标识。

8.宾词都可做主词。这从上述有关例子中可类推出。

三、标识单元的规范化

(一) 对主题词表的质疑。

主题法的标识单元必须规范化，而通过词表规范主题词，又严重地影响文献存贮和检索的效率。我们认为不借助词表又能规范标识单元的主题法，才有实用价值，才有生命力。

概念的规范化，包括统一同、近义概念的使用；限定一词多义概念在一定场合的含义；规定普遍概念的结构，使之成为一个固定概念单位。前两者在各种主题词表中占的比例都很小，后者占的比例却极大。这里首先探讨后者。

主题词表中的主题词，是一种通用性的规范概

念单位。它可以组配和揭示任何主题。要想不设词表而使所有普遍概念都成为这种通用的规范概念单位，当然是不可能的（除元词法的单元词外）。但是，概念规范化的唯一目的就是通过标引使得文献的存贮和检索达到一致性，实际上是在同一特定主题条件下的同一性。所谓概念的规范化，并不是要求标识单元必须具有通用性，而是要在同一特定主题前提下，实现文献存贮和检索中概念单位使用的前后统一。因此，如果能在每个特定主题中找到其客观存在的概念单元，并且找到划分它的规律，那么，用词表来规范普遍概念就没有必要了。

任何一个具体的文献主题，对于熟悉其专业的人员来说，都可以根据概念间的属种关系，找到与主题相应的客观固有的概念单元。如主题为《贮箱气垫增压火箭发动机的激光非破坏性检测》文献，其概念单元有"贮箱气垫增压"、"火箭"、"发动机"、"非破坏性"和"检测"。

（二）概念单元的划分方法。

概念单元是每个特定主题中最基本的完整、独立的概念单位，即标引和检索时共同遵循使用的规范主题词。概念单元受主题的制约，没有通用性。

概念单元划分的一般方法，就是从主题的复杂概念中逐层找出其本身含有的属概念。如文献主题为"低风量高风速防尘"，它本身含有的上位概念只有"防尘"。因此这一主题中的概念单元是"低风量高风速"和"防尘"。又如文献主题为"液态金属磁流体动力发电机的标准设计"，根据主题含义的层次，则分成两组概念进行划分。下面用横线表示被划分概念的各级位属概念：

液态金属磁流体动力发电机标准设计。通过属种概念的比较，这一特定主题的概念单元，则分别是"液态金属"、"磁流体"、"动力"、"发电机"、"标准"和"设计"。

概念单元是句式法最基本的语法单元和标识单元。一般地说，概念单元的内涵≥单元词，而≤主题词表中的主题词。一个概念单元可作汉语句式中的一个独立语法成分，也可以由几个概念单元组成一个复合语法成分，以表征比较专指的文献主题。如文献主题为"形变热处理对铝合金疲劳性能的影响"，则标识为：（形变）热处理影响（铝合金）疲劳性能。为使标引句式简明，复合语法成分应限制在两个等级内，因此"铝合金"、"疲劳性能"则视为一个概念单元。

（三）概念单元划分的规定。

上述划分概念单元的方法，在一般的情况下是适用的；但对于特殊的问题，需要做出如下规定：

1.概念单元必须是独立、完整的概念单位；含义不完整的修饰、限定性概念，补充意义后才能成为一个概念单元。如"长城电视机"、"C620车床"、"微电机"，应标定为"（长城牌）电视机"、"（C620型）车床"、"（微型）电机"。

2.含有肯定、否定或正反关系的概念，一律视为一个概念单元。如"标准设计"与"非标准设计"中的"标准"和"非标准"，"雷达侦察"与"反雷达侦察"中的"雷达"和"反雷达"，都分别当作一个概念单元。

3.意义不具体、不完整的属概念限制词，不能做概念单元。如"超声检测"和"互相关函数"，限制词"超"和"互"都不能单独做概念单元，"超声"、"互相关"才是一个完整的概念单元。

4.固定词组和粘合力强的复合概念，视为一个概念单元。如"物理化学"、"高超音速"为一个概念单元。

5.概念单元的划分要对应主题。如文献主题为"金属材料的热处理"，对于"热处理"来说，必然是各种金属材料，因此"金属材料"则视为一个概念单元，不再分解。

6.正确理解概念意义，切忌字面分解。如"微程序"是一个概念单元；不等于"微型"和"程序"；"轴承套圈磨床"中的"轴承套圈"是一个完整的概念单元，不能分解。

（四）同、近义概念的使用。

在特定主题中划分概念单元，不仅对一般的普遍概念进行了规范，而且也避免了因一词多义而产生的含混、歧义现象。但却没有解决同、近义概念的统一使用问题。而建专用词表来规范同、近义概念的办法，也是不可取的。

关于同、近义概念的使用问题，我们的观点是：在限制使用陈旧、冷僻和模楞两可概念的前提下，使用简化句式对同、近义概念全部标引。如不用"马达"而用"电动机"，不用"工作机"而用"金属切削机床"，不用"莱塞"而用"激光"，不用"维修"而用"维护"和"修理"。如用

钢管{裂缝 / 裂纹}焊接　　(高等)数学{《讲义》 / 《教材》}

两种句式，分别标识主题为"钢管裂缝焊接"、"高等数学讲义"有关文献。根据计算机逻辑"或"的运算方式，这种标法对计算机存贮和检索都是比较简便

的。

四、句式法的系统性

一定的检索语言形式决定着一种检索方法及其特点。汉语句式这种检索语言形式使句式法在类集和揭示主题关系方面具有一定的系统性。例如主题为"组合机床的超声检测"的有关文献，则用（组合）机床[超声]检测句式来标引。如果首先按先主词句后主述句的顺序，再以主词、述词、限词、状词的主次地位进行排列，那么就能明显地反映出有关主题的属种关系，即：机床——（组合）机床——机床检测——（组合）机床检测——（组合）机床[超声]检测。

句式法的检索语言形式，体现了分类法等级、层次的特点。完整的汉语句式——（限词）主词[状词]<适词>面词述词宾词《缀词》，根据其不同语法成分所揭示主题的等级关系，基本上可以按下列顺序排列：

主词→{(主词句)限词→缀词
(主述句)述词→限词→状词→适词→面词→宾词→缀词

汉语句式含有的这种等级、层次关系，近似于分类法中一个类系的类目关系；表示各个语法成分的概念单元，又类似于类目中的一个级位。

句式法的这种系统性不同于分类法的族性关系。它是以主题的事物主体——主词为首，来展开主题间的层次、等级关系的，不受学科门类的限制。主词相同而述词不同的主题之间关系，叫做同群主题；主词和述词都相同的主题之间关系，叫做同系主题。句式法的系统性就是这种以群系关系为基础的主题之间的等级、层次关系。如关于发电机、电动机与电机方面的主题，车床、镗床、铣床、磨床与机床方面的主题，从分类法的角度看，前者都应从属于后者。句式法则不考虑前后者的属种关系，而将前后者都分别作主词，组成各自的独立群系主题关系。

与叙词法和分类法比较，句式法这种群系关系所揭示的主题范围则具体、集中、明确，更能满足读者的实际检索需要。

五、计算机检索

能否实现计算机快速、准确地检索，是对句式法的严峻考验。去年，我们进行了一次句式法微机检索演示（河北省教委批准的课题），对近千种科技图书做了实际标引、建库和程序设计。通过这次机检实验，初步证实了句式法是一种完全适于机检、并达到预期目的的主题法。这次的检索时间，平均一种主题为半分钟，估计存贮5万种以下图书资料，检索时间为2～3分钟，最多不超过5分钟。

这里要强调的是：计算机检索的前提条件，要做到文献存贮时的主题全面、正确的标引（多种句式运用，同、近概念使用，综合标引与分析标引，上位主题标引等），以防漏检；要有合理的建库方案；准确而周密的编程。

（一）机检步骤。

1.正确理解主题，确定概念单元。要检索的有关文献主题分别为"在高温条件下，盐对钢的腐蚀"、"齿轮的径向跳动测量"、"模具钢的充氧压力铸造"。根据其不同主题的含义，各主题中的概念单元确定为：高温，盐、钢、腐蚀；齿轮，径向跳动，测量；模具，钢，充氧，压力，铸造。

2.按照主题，选标一种句式。上述各主题，在存贮标引中可能用几种句式标识，但检索时只任意标出一种句式即可（如果遇到同、近义概念时，也只标出其中一个概念）；

(1) 盐<高温>腐蚀钢

(2) 齿轮径向跳动测量

(3) (模具) 钢[(充氧) 压力]铸造

3.填写机检单。将上述检索标引句式中的概念单元，按其语法成分填入机检单内。限词一般是1～3个，表格应留有位置。

计算机检索单

词类 / 句式	主词	述词	限词	状词	适词	面词	宾词	缀词
(1)	盐	腐蚀			高温		钢	
(2)	齿轮	测量				径向跳动		
(3)	钢	铸造	模具	(充氧)压力				

中国图书馆学报(季刊)1992 年第 2 期
ZHONGGUO TUSHUGUANXUEBAO

●孙清兰

高频词与低频词的界分及词频估算法

齐夫第二定律揭示了低频词的分布规律,给出:

$$I_n / I_1 = 2 / n(n+1) \cdots\cdots \quad (1)$$

式中，I_n 代表文中出现 n 次的词汇数量。比值与文章长度无关〔1〕。

高频词与低频词分界有个临界值，这是 Dono hue, J · C · 于 1973 年提出的〔2〕。其计算公式为:

$$n = \frac{1}{2}(-1+\sqrt{1+8I_1}) \cdots\cdots \quad (2)$$

可见，公式 (2) 依赖于文中出现一次的词数。

本文也给出以下一条高频词、低频词分界临界值的计算公式:

$$n = \frac{1}{2}(-1+\sqrt{1-4D}) \cdots\cdots \quad (3)$$

公式 (3) 与公式 (2) 的不同点是，(3) 式取决于文章中的不同词数 D，而 D 显然比 I_1 容易得到。进一步，我们可以给出公式 (3) 的以下简化形式:

$$n = \sqrt{D} \cdots\cdots \quad (4)$$

利用这一简化公式界分高频词、低频词，不仅与 Dono hue 公式的效果一致，而且计算简捷，使用方便。

本文通过验证，得到了不同数量同频词的词频估计公式，从而进一步丰富了词频分布规律的内容。

一、高频词与低频词的界分方法

(一) 词的等级的确定。

在研究词频的分布规律时，对词的等级有几种不同的确定方法。我们这里采用的是最大值法，即把文章中的词按出现次数由高至低顺序排列，遇到同频词任意排序。词的等级取作同频词词序的最大值。显然，不伴有同频词的词 (同频词总数为 1)，其等级就是它的词序值。

表 1 以文献〔3〕为依据。其中，文章长度 T=3907，不同词数 D=813。

表 2 以文献〔4〕为依据。其中，总词数 T=6958，不同词数 D=1170。

(二) 高频词与低频词的界分公式。

从表 1、表 2 可见，最后一个词的等级恰好等

(二) 含有概念排除的文献主题检索。

实际检索中，有时读者可能需要查找除某一个或几个方面之外的某主题文献。如检索主题为“除喷雾淬火之外的各类模具的各种淬火”、“除金钢镗床和非标准设计之外的其它各类镗床的各种设计”的有关文献。这种要排除的概念，计算机工作时，是逻辑非运算。进行检索句式标引，必须要用逻辑非符号表征出它的特殊含义。上面两个主题用句式标识为:

1.模具[喷雾]淬火

2. (金钢) 镗床[非标准]设计

填写机检单时，要将排除概念连同它的逻辑非符号一并写入相应词类项。

综上所述，句式法的优点有以下几个方面:

1.以汉语为基础的检索语言形式，表意直观、明确，便于掌握，能准确标识各种错综复杂的文献主题，能进行多途径检索。

2.具有专指性和一定的系统性检索功能。

3.使用语法职能符号，保证了检索的准确性，避免了检索中的“噪音”。

4.提高了主题的存贮和检索效率：概念单元取代了词表，极大地减轻了存、检的工作量；标有语法职能符号的标识单元，有效地节省了计算机运行时间。

5.不仅适于联机检索，而且能实现微机的独立存贮和检索。

(来稿时间：1991.6.编发者：刘喜申。)

An Interpretation of Library Management / Zhao Chengshan / / Bulletin of the Library Science in China / China Society of the Library Science. -1992, 18(2).69～71

The definition of library management can be summed up as follows: It is a kind of activities conducted by the library administrative personnel for the purpose of achieving the optimun goal of a certain activity of library systems, who, by way of several processes of planning, organizing, controlling, etc., optimizes the combination of all the library resources to reach the hight of improving the social benefit of library systems. 6 references.

Theory of systems —— Applications
Library management —— Studies
Library undertaking —— organization and management
G251

Time Management and the Director of the Library / Chen Shu / / Bulletin of the Library Science in China / China Society of the Library Science. -1992, 18(2).71～73

Accompanied with the radical increase of information and wealth of man, the library world is now confronted with the challenge of time. A director of the library will consequently be far behind if he handles his work merely by the virtue of his good intentions and experiences. He has to know well the theory and technique of time management, besides. The chief items of time management are: the 4D law, the exception law, the negative law, etc. as well as those techniques derived from these laws. In short, time management is none other than the knowledge of approaching ways on eliminating waste of time, thereby, it will be able to attain the goal of running a library. So far as the significance of the modern management is concerned, the wing for the Chinese libraries to soar is exactly the time. 2 illus. 6 references.

Library undertaking —— Scientific management
Directors of libraries —— Personal qualities
Time management —— Theories
G251-36

A Preliminary Probe into Subject Indexing Retrieval of Chinese Scientific and Technical Documents / Subject Indexing Research Group, Library of Hebei Mechanical and Electrical College / / Bulletin of the Library Science in China / China Society of the Library Science. -1992, 18(2).74～78

The abbreviation of Chinese sentence-mode subject indexing is the "sentence-mode method" — a new method for Chinese scientific and technical document subject indexing retrieval. The Chinese sentence-mode being a form of the retrieval language, is compatible with some characteristics of subject indexing and classification. The article also makes an approach to a new may for the standardization of mark unit. In each particular subject, there exists objectively a kind of "concept unit" which one may follow to use. The "concept unit" is not like the unit word of the basic word method, nor is it like the subject word derived from a thesaurus artificially standardized. It is an objective, intrinsic concept unit separated out from a particular subject, i.e. a kind of standard subject word of a special form without a thesaurus. The method has already been retrieved and tested by a computer. 1 table.

Subject word method —— Approaches
Sentence-mode method —— Reviews
G254.2

The Demarcation of High - and Low - frcquency Terms and Ways of Estimating the Frequency of Terms / Sun Qinglan / / Bulletin of the Library Science in China / China Society of the Library Science. -