中国图书馆学报(双月刊)1996 年第 5 期
ZHONGGUO TUSHUGUANXUEBAO

●贾同兴

信息管理中的不确定性研究

ABSTRACT Having put forward both the different interpretations in the natural language of scientific and technological Chinese and the manifold indeterminate research subjects existed in the multimedia management, the author makes an investgation on the caux, types, causes sox the occurrence and measures for removing them. 17 refs.

SUBJECT TERMS Multimedia technology—Studies Scientific and technological languages—Indeterminacy
CLASS NUMBER G350

随着计算机技术、通信技术的发展,计算机处理的信息介质就不再仅限于人工语言、自然语言,也包括多媒体信息载体介质。在信息载体范围扩大的同时,信息管理中的不确定性因素大大增加,它严重影响着信息的获取、表达和使用。因此,对不确定性进行研究,在更高理性层次上认识它,以加强信息管理建设和指导信息管理系统的设计开发,是一项带战略性的任务。

1 信息管理中不确定性的性质和类型

1.1 性质

信息管理中的不确定性,首先是因为事物本身的复杂性和表露不充分以及人类认识上的局限性(包括知识、知识结构、经历、文化背景及思维方式等)引起人们对事物类属边界和性态认识上的差异;其次是因为信息管理是多种人参与管理多种人的思维产品,是受强烈人文因素影响的事业;第三是因为信息管理系统是一个开放的、动态的巨系统,其动力学性质异常复杂多变。信息管理中的不确定性实质上就是人类对信息认识上的类属不清、性态不明的模糊性。

1.2 类型

对于信息管理中的不确定性,国外虽有不少文献报导[1~3],但研究的视野太狭小,仅仅局限在情报检索语言的人工语言上。狭小的研究领域大大局限了人们更高的理性思考,影响了人们对信息管理中不确定性性质的认识,也影响信息管理系统的设计开发。

从对自然语言不确定性的研究中,我们认识到不确定性是信息的一种重要属性,是一种客观存在,也是一种推动认识深化的力量。为管理上的方便,我们不得不在一定的阶段对信息的这种属性加以限制,施以约束,相对的求精、求准和求确定性,以便集中我们思维的焦点,阶段性地解决问题。为此我们可以把不确定性类型图示出来(见图 1)。

2 不确定性的研究历史

关于信息管理中不确定性的研究,过去主要集中在检索语言的人工语言,而在人工

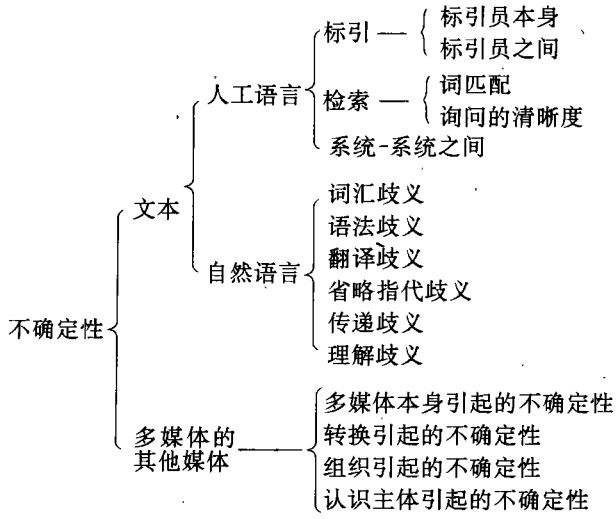

图1 不确定性类型

语言中又主要集中在标引、检索及系统方面。

由于认识水平局限和计算机技术水平限制，人类对人工检索语言的研究主要建立在经典集合论的基础上，尤其是建立在“非A即B”的二值逻辑上，认为不确定性是由主题领域的性质(硬科学和软科学)及标引员培训水平、经验和标引词表的性质不同造成。因此为了减少不确定性，在“标引为主理论”指导下，主要只对标引员自身、标引员之间及检索的不确定性(**Inconsistency**)进行了研究[4]。文献[5]则用控制论的理论分析了标引和检索中的不确定性，总结出三个原则：(1) 不确定原则(主题标引是不确定性的和概念性的)；(2) 不变化的原则(检索询问的变化应等同文献标引)；(3) 复杂性原则(探索过程是一种神秘莫测的过程)。并针对现行联机目录式检索提出了一个标引模式：一个终端用户的词表；一个检索提问编制前端机；一个命令语言和一个自然语言接口。文献作者认为这种模式可以降低标引和检索中的不确定性。近年来，信息管理中不确定性的研究领域有所扩展，提出了情报检索系统之间也存在不确定性，认为“由于联机存贮和检索系统的普遍化，很多系统(或数据库)的文献集合重叠，然而不同系统使用自己的标引词表、标引规则，产生了不同系统间的不一致性”[6]。文献[7]以药物学为主题，对一些该主题的系统进行了分析实验，发现在药物配方、药物动力学和其他药物学方面概念处理、标引方法有相当大的不一致性，并认为这种不一致性是由于使用不同的标引词表，采用不同的标引规则和由不同标引员进行标引造成的。

文献[8]对消除标引和检索中不确定性的对策进行了深入的研究，认为标引员自身的不确定性是学习和获取知识的过程；标引员之间的不确定性是标引员在标引的主题领域和分类知识上的差异；系统间的不确定性是分类模式的差异。在以上分析的基础上提出了“以变治变的元知识结构”方法，即把不同标引员用的词表、不同的分类表和不同系统都容于一个综合的知识结构——元知识结构中，并建立彼此的联系，允许不同的部件间互相通信，通过这种提供包容变化以消除变化的不确定性。

该文献还提出了智能原则，即依据标引人员的知识结构，运用框架和神经网络结构创制一种推理部件，并通过积极的计算机会议技术促进交互，帮助检索者在语义层次上优选最相关的概念(包括关键词、**FOLDER**和组织)增强推理能力，帮助搜索者决策。文献[9]在以上两原则的基础上提出了基于知识的亚利桑大学分析家情报系统(**AAIS**)，缓解了信息管理中的不确定性。

3 自然语言信息管理中的不确定性

人类的自然语言是一个极为复杂的符号系统，是理解人类心灵和文化的关键，是最高效、最方便的信息载体。人类接收信息的80%以上要靠文本。

专家预测，21世纪将是以智能研究为核

心的世纪，我国语言学家范继淹认为："自然语言理解是人工智能的核心"。所以 1990 年世界第 13 届计算机语言学大会把"处理大规模真实文本"定为今后的战略目标。自然语言的研究最富吸引力、挑战性和困难性，其最大的困难就是自然语言中的歧义。在过去几十年的研究中，专家们在对人工智能、机器翻译、自然语言理解及情报检索系统接口的研究中形成共识，认为"歧义是自然语言处理的第一困难问题"[10]。当然也是自然语言信息管理中第一困难问题，其困难主要来自自然语言的语音、语法、语义和语用各层面上存在的歧义现象。

所谓歧义，就是人们对知识① 认识上的模糊性，包括二义性、不精确性、不完整性和不准确性以及一词的多种理解。过去对歧义的研究，多集中在纯语法研究上。在机器翻译、自然语言理解和自然语言接口的研究中虽也做了一定工作，但专门针对情报检索的研究则很少。

对歧义定义的研究尚未形成共识。有的定义为"有多种理解或多种意义"；有的定义为"一个表层结构的句子对应两个或两个以上的深层结构"；还有的定义为"一个自然语言的词串，经过自然语言的文法分析器后，会产生多个符合语法的输出"。这些定义有的显然是从语言的一个层面上的局部定义，有的则太笼统，从而使研究工作开展不充分、不深入，难以给出实质性定义。为了讨论的可行性，笔者试提出以下定义："歧义是认识上有差异的人们对自然语言固有的、模糊性的模糊认识"。人们由于认识、知识结构、文化背景、培养和经历、思维方式等的差异，对本来无歧义的事情都可能形成认识上的歧义，加之自然语言本身固有的模糊性、灵活性、不精确性、不完整性、二义性、无限性等，在两者匹配时，可能产生模糊认识。

对歧义的分类涵盖面太窄。有的分成"词汇歧义、结构歧义和语话歧义"[12]；有的分成"传统语法发现的歧义，结构主义发现的歧义和转换生成语法发现的歧义[13]。笔者则把歧义分成词汇歧义、语法歧义、翻译歧义、指代省略歧义、传递歧义和理解歧义。

词汇歧义。这是最主要、最大量的歧义，如 **The pig is in the pen**。这种歧义主要因为多义词引起，各种词类都有这种歧义。

句法歧义。如"咬死/猎人的狗"与"咬死猎人/的狗；**They are flying planes** 与 **They are flying planes**。句法歧义是指句中无歧义词而产生的歧义。

翻译歧义：一种本无歧义的语言，经翻译后却产生了歧义。如汉语成语"眼不见，心不烦"，从英文"**out of sight out of mind**"译成俄文，再译回英文检验时就成了"又瞎又疯"。

指代省略歧义。指代省略隐喻也往往造成歧义。如 **He said that yesterday he had finished the work** 与 **He said yesterday he had finished work**。第二句因为省略了 **that**，所以 **yesterday** 既可以修饰 **had finished**，也可以修饰 **said**，既可以译成"他说他昨天就已完成了那个任务"，也可以译成"他昨天说他已完成了那个任务"。

传递歧义：信息在传递中发生畸变、增值或丢掉信息。这种情况多见于传闻中，道听途说，以讹传讹。

理解歧义，即对同一事物有不同理解。这是一种高于词和句级别的更高层更容易引起不确定性的歧义类型。

歧义的产生实质上由两方面引起：一是自然语言固有的现象，如一词多义，词的兼类；二是由人的认识不同而引起的理解差异以及所采用的处理方法不同所引起的歧义。

因为汉语的特点产生了一些特有的歧义[13~15]，如汉语书面无法表达语气、语调、读

① 知识＝客观事实＋主观信息＋主观信息与客观事物的一致性关系[11]。

音、轻重音、停顿引起的歧义;汉语词间无空格,分词时引起歧义;古代与现代汉语之间的歧义;短语、词组的边界难于确定,如介词短语的右界、方位结构的左界;双向动词和三向动词+结构助词引起的歧义,如"喜欢的",既可指施事,也可以指受事,像"喜欢的举手"与"你挑喜欢的";因汉语中词组类型与句法功能不存在一一对应造成歧义,如 **NP+VP** 可以是主谓,也可以是偏正,如"计算机设计";**VP+NP** 可以述宾,也可以偏正,如"出租汽车"。以上列举了汉语特有的歧义,当然还有与其他自然语言共同的歧义。

汉语的歧义研究,尤其是科技汉语的研究急待加强,以适应大规模文本处理信息的需要。

关于消除、限定自然语言中歧义的对策,有人提出利用语境上下文、变换句式、调整语序和增词。有人提出短语分析法,利用上下文和利用规则;还有人提出优选法,深层语义分析法和语境分析法等。我认为:消除歧义最重要的方法是限定系统的专业领域,首先限制在科学技术领域内,再进一步限定某一专业甚至专业的一个小类内。这样有些歧义不用定义而自然消除,如"花",在植物学专业领域是无歧义的。其次是"以变治变",即在系统内设有大的知识库、语料库、规则等尽可能包罗自然语言中的歧义及对策。以上两条件也是未来智能系统的要求。利用子语言消除歧义也是一种极为有效的措施。

用户是系统运行的专业领域的专家,他们是知识库建设的重要来源,系统遇到歧义时应请教专家用户。利用人机交互也是系统建设的全过程必须注意使用的、有效的技术手段,因为它把静态的知识与动态的知识结合了起来。

一种语言系统里的错综复杂现象和精妙之处,往往在歧义现象中反映出来。因此,歧义的研究不仅对信息管理、人工智能、自然语言理解,而且对人脑的语言和思维机制的研究都非常有意义。

4 多媒体管理中的不确定性

多年来,人与计算机的交互一直局限于文本,这与人类自然形态中使用多种信息媒体相比,严重影响了人类本来所具有的通信能力。多媒体的出现增加了人机交互的手段。

多媒体的信息管理不仅是数据库的增删、插改、检索等常规操作,而且更多的还有识别、理解、转换、关联、压缩、合成、创作等特殊操作,极大地增加了系统的功能,也增加了系统的不确定性。这些不确定性包括:媒体自身的不确定性;媒体转换不确定性;主体认识上的不确定性;扫描引发的不确定性。

(1) 媒体本身的不确定性。多媒体也是一种媒体的不确定性,在图形、图象、声音中,似乎给人留下更大的联想空间,可能产生更多的不确定性。

(2) 媒体转换所产生的不确定性。声音与文字的互相转换会产生歧义。如将话语"我想起来了"变成文字,语调、语气、读音和轻重音信息就不得不丢掉,由于语气、重音的无从区别从而造成歧义。

(3) 主体认识上的不确定性。这种不确定性是由认识主体的知识、知识结构、经验和经历、培养水平、文化背景和思维方式等的差异引起。

(4) 扫描引发的不确定性。多媒体或超文本是用语义网组织知识的,当人们在网上进行联想扫描时,受网上感兴趣信息的影响,顺着用户知识结构(即知识、心理、经历、习惯、情绪等个性动态)无固定路线和无固定方向的自然流动,一旦远离原点就会产生迷失方向的情况,或不知自己目前在网上的位置;或虽知道网上有某些位置但不知如何到达;或忘记从分支返回和忘记原先计划要到达的位置;或不知道网上是否还有相关的结点;或忘记已访问过的结点等等[17]。(下转第 27 页)

馆被合并入上海图书馆，史称“四馆合并”，1960年又把市少年儿童图书馆也并入（1987年市少年儿童图书馆恢复了独立建制）。上海图书馆从此成为一家藏书很丰富的综合性的大型公共图书馆了。可是，历史证明，这一熔化特色铸造大综合的政府行为是不够明智的，特色被淹没了，而大综合运转不灵，无法从管理、从服务、从科研、从交流上走入与它身份相称的境地。象上海这样一个大城市的公共图书馆事业，如能以综合的上海图书馆为托依，以上述的市级特色图书馆（或更多的市级特色图书馆）为卫星，再以具有众多特色服务手段的区、县、街道、乡镇图书馆为网络，则它将从整体上达到既具有核心馆又具有特色馆和特色服务纷呈的基层馆这样一个大都市的强有力的三级公共图书馆事业网。可惜，这仅仅是马后炮。

特色服务是读者工作的深化。它是从传统的借借还还的老框之中脱颖而出的具有时代特征、符合读者需求的新的服务内容。它是公共图书馆转变服务观念，摆脱服务困境、提高服务质量的方向，是公共图书馆从“门前冷落车马稀”的苦境和“文化的沙漠”走入阳光大道的大门。当一个馆的特色服务达到成熟程度，在条件允许情况下，可转建成特色馆。如上海虹口区曲阳图书馆影视资料的收藏和服务很有特色，但因为它还兼容地区综合性服务工作，所以门口只好同时挂着两块招牌：“曲阳图书馆”和“影视资料图书馆”。

黄恩祝 上海图书馆研究馆员，已退休。通讯地址：上海市真如西村118号402室，邮编200333。

（来稿时间：1995—01—29。编发者：刘喜申。）

（上接第17页） 总之，对信息管理中的不确定性，尤其是自然语言及多媒体管理中的不确定性急需进行深入研究，其中汉语科技文本及自然语言的歧义问题应是近期研究的重点，它对解决主要信息载体——文本处理极为重要。

参考文献

1，4 Blair. D. C. Indeterminacy in the Subject Access to Documments. IP&M，1986，22(2)229～241

2，5，8， Bates，M・J. Subject Access in Online Catalogs：A Design Model. JASIS，1986，37(6)：357～376

3，7 Barber，J.，Moffat，S.，Wood，F. Cases Studies of the Indexing and Retrieval of Pnarmacology Papers. IP&M，1988，24(2)：140～150

6 White，H. D.，Griffith，B. C. Quility of Indexing in Online Data Base. IP&M，1987，23(3)：211～224

9 Chen，H. Danowitz，A. K.，Mchenry，W. K. Explaining and Alleviating Information Management Indeterminizm：A Knowledge－Based Framwork. IP&M，1994，30(2)：557～577

10 Yickery，B.，Vickery，A. Online Search Interface Design，Journal of Docnmentaon. 1993，49(2)：102～187

11 金雅芬. 关于知识研究的历史成果之批判与继承. 计算机科学，1995，22(3)：21

12 张克理. 英语歧义结构. 天津：南开大学出版社，1993：1～2

13 吴蔚天. 汉语计算机语言学——汉语形式语法和形式分析. 北京：电子工业出版社，1994

14 周明等. 用人机对话解决歧义问题. 情报学报，1991，10(4)：273～281

15 刘开瑛等. 自然语言处理. 北京：科学出版社，1991.

16 陆俭明. 八十年代中的语法研究. 北京：商务印书馆，1993.

17 Foss，C. L. Tools for Reading and Browsing Hypertaxt，IP&M，1989，25(4)：407～418

贾同兴 河北大学信管系副教授。通讯地址：保定市欧和岭1号，邮码071002。

（来稿时间：1996—04—22。编发者：翟凤岐。）