MIS 智能处理的近似评判法及其算法研究*)

谈文蓉　杨宪泽
（西南民族大学计算机科学与技术学院　成都610041）

摘　要　在MIS的设计中，智能技术的使用是一大趋势[1,2]。本文围绕关键词的智能检索问题，完成了三部分工作：1）探讨了近似评判方法在MIS智能处理中的应用；2）给出了一个文献检索智能接口的设计；3）提出了相应的规则索引算法。

关键词　MIS，文献组织，近似评判，智能接口，规则索引

Analogic Judgement Methods Regarding Intelligent Processing of MIS and its Algorithm

TAN Wen-Rong　YANG Xian-Ze
(College of Computer Science and Technology, Southwest University for Nationalities, Chengdu 610041)

Abstract　MIS design features the application of intelligent technology, which is a development tendency at present time. Focusing on intelligent index based on keywords, this paper comprises three parts. the first part discusses about the application of analogic judgement regarding MIS intelligent processing; the second part develops an intelligent interface design of document index and the third part proposes the corresponding rules index alogrithm.

Keywords　MIS, Literature organization, Analogic judgement, Intelligent interface, Rules index

管理信息系统(MIS)是由人、计算机等组成的能进行信息的收集、传递、存储、加工、维护和使用的系统它综合了计算机科学、经济管理理论、运筹学，统计学而形成为一门系统性边缘性学科，也是一门国内外到目前为止尚不很完善的多元目的新兴学科[3]。

从计算机应用角度看，MIS涉及的行业很多，它在社会经济活动中也起着越来越重要的作用，因此，它的理论和方法的研究都具有十分重要的意义。针对MIS涉及的数据种类多、数据量大且数据分散的情况，我们在MIS的智能处理方面进行了一些有益的探索，文[4]探讨了MIS的检索算法和纠错接口，本文则在此基础上探讨了近似评判方法在MIS智能处理中的应用并给出了相应的算法。

1　近似评判方法

1.1　构思

文中涉及的科技成果记载了职工已鉴定或评审的项目、专利，公开出版的专著、编著、译著，国内外正式期刊上发表的论文。科技成果的评奖是一件复杂的事情，要保证公正就必需有定量的操作方法，尽可能减少人为因素。我们的作法是有一部分直接由计算机计分，比如项目来源：国家级10分；省部级8分；厅局级6分；自选4分。这些可直接计分的包括如推广项目经济效益，鉴定意见的水平，发表论文的杂志级别，单篇与系列论文等等。

这些直接计分项越多，越可操作。此外，这还有一大用途，比如，这届评奖预计不超出30项，但上报150项，通过计算机录入、排序可劝排在后面的自动放弃，这避免了大工作量。

对于不能直接计分的，实际上是性能指标量度“模糊”，如社会效益、技术难度、重要性等等。假如我们用精确的方法来评判这些性能，则是一件困难的事情，尤其是精确的综合评判就更加困难了。然而，如果运用模糊集的概念，这些问题可以基本得到满意的解决。

1.2　近似评判

对这些模糊指标进行评判，采用评委填表的方式进行。即：

其它指标	重要性	技术难度	社会效益
评委意见			

让他们对表格重要性选重要、比较重要、一般、不重要四种答案中挑一种填入；而技术难度选难、较难、一般、不难；社会效益选好、较好、一般、不好。

下面是对某一成果的评判：

当对评委表进行统计时，对重要性有50%认为重要，40%认为较重要，10%认为一般，没有人认为不重要，那么这个单因素评判结果的模糊集 $R1=(0.5\ 0.4\ 0.1)$ 同样对技术难度评判的模糊集 $R2=(0.4\ 0.3\ 0.2\ 0.1)$，对社会效益评判的模糊集 $R3=(0\ 0.1\ 0.1\ 0.6)$. 由 $R1$、$R2$和 $R3$构成矩阵：

$$\underset{\sim}{R}=\begin{bmatrix}0.5 & 0.4 & 0.1 & 0\\ 0.4 & 0.3 & 0.2 & 0.1\\ 0 & 0.1 & 0.3 & 0.6\end{bmatrix}$$

但由于成果不同，他们对各种性能的要求也不尽相同。比如，开发项目可能对技术难度要求较高，而理论研究可能要求重要性更高。假如，有某一成果，他们对重要性的要求为50%，对技术难度的要求为20%，对社会效益要求30%，则构成的模糊集为：

$$\underset{\sim}{A}=(0.5\ 0.2\ 0.3)$$

*)基金项目：国家民委重点科研项目基金(编号234408)资助项目，西南民大重点资助项目(No. 04NZ003)。谈文蓉　副教授，研究方向为计算机网络、多媒体技术和管理信息系统。杨宪泽　教授，研究方向为自然语言处理、算法与数据结构。

因此，对这一项目的近似评判结果为：

$$B=A\cdot R=(0.5\ 0.2\ 0.3)\begin{bmatrix}0.5 & 0.4 & 0.1 & 0\\ 0.4 & 0.3 & 0.2 & 0.1\\ 0 & 0.1 & 0.3 & 0.6\end{bmatrix}=(0.5\ 0.4\ 0.3\ 0.3)$$

将近似评判结果归一化，在这里0.5+0.4+0.3+0.3=1.5，则将近似评判结果改为：

$$\left|\frac{0.5}{1.5}\ \frac{0.4}{1.5}\ \frac{0.3}{1.5}\ \frac{0.3}{1.5}\right|=(0.33\ 0.27\ 0.2\ 0.2)$$

最后得到三项分数(第一项乘100，第二项乘80，第三项乘60，第四项乘40)：

33+21.6+12+8=75.6(分)

2 智能接口

本节的工作提出了一种自然语言检索接口方法，其背景是：用户对MIS检索要求越来越高，他们希望系统对各种方式的提问都能有满意的回答(输出需要的结果)。对于该复杂的提问形式，要么训练检索操作人员(并依据他们积累的经验)以适应不同环境下的要求；要么让MIS在检索功能方面具有求解复杂问题的能力。本节的工作在于后者，为MIS设计了一个子模块，能理解一般检索用语。这是一种自然语言检索接口的方法，它提高了自动检索能力。

计算机如何理解汉语，是我国人工智能研究者探究的一个核心问题，难度极大，涉及到一系列语言学和计算机的理论问题。首先面临的汉语自动分词就是目前公认的难题。基于此考虑到有限范围内可实用原则，把一般检索用语作为汉语集合中一个很小的子集来考虑，且在原文分析(输入的检索要求语句)，原文转换(转换成计算机能识别的操作任务)阶段，力求简化，从而比较成功地满足了检索要求。

2.1 结构格式分析法

这种方法来源于自动翻译，它根据词的分布状态把词分成若干类别，然后根据所分出来的这些词的类别确定词组类型，把语言的各个词组类型构成一份词组类型清单。在进行自动翻译时，首先借助于机器词典把词的类别特征记录在文句中所有的词上，然后，计算机把所要分析的文句与词组类型清单进行比较，从而在文句中找出相应的词组类型。这样，就可确定文句中词与词之间的句法关系。词组类型也就是一种结构或格式，因此把这种方法叫做结构格式分析法。这实质上是一种模式匹配的方法，本节做了改进然后引用。

2.2 自然语言检索接口使用的语言库

关键词库是原有的，检索围绕关键词进行。关键词包括文献主题词、作者、单位、学科和获奖情况等。

检索用语库存储足够数量的单词，它们与关键词组合，最后将转换成计算机能识别的操作任务。

例如，"检索数据结构和人工智能文献"

这句中，"检索"，"和"，"文献"这3个单词存储在检索用语库中。单词在子模块逐渐完善中不断增添和修改。语法规则库是存放检索用语规则，不合理的规则易于修改或删除，新规则可以添加。规则以产生式表达。

Ri：IF RLS THEN RRS ($i=1,2,\cdots,n$)

其中，Ri 为规则集中第 i 条规则；RLS 是第 i 条规则条件部分；RRS 是第 i 条规则结论部分，是计算机主要完成的检索操作。

条件部分是以结构格式分析法实现的检索用语助记符，它们表示了一定的语法关系。

例如：

条件部分为：Vnonn1

检索	数据结构	和	人工智能	文献
v	n	o	n	n1

检索	计算机系	1990年	获奖	文献
v	n	d	n	n1

条件部分为：Vndnn1

2.3 原文分析

原文分析系指对检索用语分析，包括自动分词和单词助记符标注两部分，步骤如下：

步骤1：一条检索用语分划成单一字符 $X1,X2,\cdots,Xm1,X2,\cdots,Xm1,X2,\cdots,Xm$

步骤2：决定关键词中一个关键词最大字符长度 $Lmax$，最小字符长度 $Lmin$；检索用语库最大字符长度 $Ymax$，最小字符长度 $Ymin$(这实际上是分词前预知的)。

步骤3：逆向匹配，取检索用语句最后的 $Lmin$ 个字查关键词库。若查不到，加入一个字重复此工作，直至字符数为 $Lmax$ 为止。min 个字查关键词库。若查不到，加入一个字重复此工作，直至字符数为 $Lmax$ 为止。

步骤4：若实施步骤3查不到关键词，去掉语句中最后一个字，再实施步骤3，直至整个语句只剩下 $Lmin$ 为止。

步骤5：若实施步骤3直到关键词 $Xi,Xi+1,\cdots$ 则将 Xi，$Xi+1,\cdots$ 标注 n。且 $Xi,Xi+1,\cdots$送一专用词存储区，n 送一专用助记符存储区，此外，剩余 $X1,X2,\cdots,Xi-1$再实施步骤3。

步骤6：语句剩余部分查检索用语库，基本形式如步骤3～步骤5，但其中 $Lmax$ 这里为 $Ymax$，$Lmin$ 这里为 $Ymin$。若查到检索用语，不仅要送助记符到存储区(检索用语的助记符事先已规定好)，而且依据 $X1X2\cdots Xm$ 的下标序自动调整助记符位置。

到此为止，原文分析阶段结束。这里要注意的是，我们不承认用户输入的不精练检索用语，因此，系统将去掉冗余信息。

例如：

请检索	数据结构	方面的	文献
v	n		n1

原文分析阶段结束时，只留下Vnn1和检索数据结构文献，而"请"和"方面的"无论在关键词库和检索用语库中都查不到，就自动去掉了。

2.4 原文转换

原文转换将利用助记符存储区存储的助记符串查语法规则库，与规划匹配，找到相应规则后，据规则结论部分要求，系统将转入相应程序段，再依据词存储区的关键词完成用户检索要求。

3 规则索引算法

3.1 索引原则

索引是一种映射关系。本文的工作希望满足：

(1) 能把较大的规则(条件部分)分配到某一特定的地址范围。

(2) 能把 N 条规则较均匀地分配到不大于 $2N$ 的不同存储单元。

(3)索引算法所花的代价较小。

3.2 索引算法思路

(1)每条规则条件部分可视为一个字符串 Bi，含有单个

字符 $Bi1,Bi2,\cdots,Bik$ (k 为字符串长度)，可分划

$LEN(Bi\$)$；求字符串长度

do j from 1 to k

$Bi\$j \leftarrow MID\$(Bi\$,j,1)$；切分字符串成单一字符。

(2)在计算机内，每个字符可转换成 ASCII 码介于0～225 之间的十进制数。

$Ci1=ASC(B\$i1),Ci2=ASC(B\$i2),\cdots,Cik=ASC(B\$ik)Ci=Ci1+Ci2+\cdots+Cik$ $(i=1,2,\cdots,N)$

到此为止，规则 Ri 对应 $Ci(i=1,2,\cdots,N)$。$C1,C2,\cdots,Cn$ 中有的数可能相同，但由于规则条件含有字符较多，引起 $Ci=Cj$ $(i\neq j)$的条件为：

①字符串单个字符都相同(例如，*abcd bacd* 等)

②$Ci1+Ci2\cdots+Cik=Cj1+Cj2+\cdots+Cjk$。这种情况出现的频率不可能太高，而且我们允许有一定的相同，只要相同数的个数$<<N$。

3.3 算法描述

A1：设有 N 条规则，规则 $Ri(i=1,2,\cdots,N)$条件部分视为字符串 Bi，可分划成单个字符 $Bi1,Bi2,\cdots,Bik$ (k 为字符串长度)。

A2：(在计算机内，每个字符可能转换成 ASCII 码介于0～255之间的十进制数)，求这一转换：

$Ci1=ASC(B\$i1),Ci2=ASC(B\$i2),\cdots,Cik=ASC(B\$ik)$。

A3：$Ci=Ci1+Ci2+\cdots+Cik$ $(i=1,2,\cdots,N)$。

A4：扫描一遍已转换成数值的集合 C，求 $Cmax,Cmin$ ($Cmax$ 和 $Cmin$ 为集合 C 的最大值和最小值)。

A5：变换 $M=INT(1/a(Ci-Cmin))$(式中 a 为大于或等于1的正整数，$i=1,2,\cdots,N$)，

A6：开辟一个数组 F，容量为 $INT(1/a(Cmax-Cmin))$，建立规则 Ri 对应 $F(Mi)$索引 $(i=1,2,\cdots,N)$，若有规则 Ri 和 Rj $(i\neq j)$ 使 $F(Mi)=F(Mj)$，采用链接方式将它们链接。

3.4 匹配操作

匹配操作时，有关事实(假定词组、词组特征等)仍视为字符串，利用一专用程序进行如下步骤：

步骤1：将事实字符串分划成单个字符。

步骤2：求 $Ci1,Ci2,\cdots,Cik$ 和 Ci。

步骤3：计算 $Mi=INT(1/a(Ci-Cmin))$

步骤4：让 Mi 对应 $F(Mi)$的地址，这时若无规则，失败退出，若为唯一规则，成功，若为链，按一般方法继续链上的匹配操作。

3.5 算法分析

索引算法旨在提高系统运行速度，其效率可按编译原理[5]进行分析：一般的规则匹配实际上是字符串的匹配操作，切分成单个字符与索引匹配的方法是共同的。不同的部分是：一般匹配将进行两个字符串单个字符的一一比较和传送，索引匹配有求字符串长度，求 Ci 和 Mi 的操作，按文[6]类似的指令流方式计算，这两种匹配方式指令执行时间相差无几。但是，一般匹配成功最大次数为 N 次(N 为规则数)，最小次数1，平均次数1/2(1+N)；而索引匹配成功最大次数为 j 次(j 为 $C1,C2,\cdots,Cn$ 相同数的个数，$j\ll N$)最小次数1，平均次数1/2(1+j)。

说明：(1)在规则 Ri 中，可能出现 AND 条件和 OR 条件，由于所有与条件满足才可以为规则匹配，因此所有 AND 条件作为一个字符串最后变换成 Mi；对于 OR 条件，由于只需要其中一个条件满足即可认为规则已匹配，所以每一 OR 条件单独作一个字符串变换成 $Mi1,Mi2,\cdots,Mid$ (d 为 OR 条件个数)，采用链接方式。

(2) $Mi=INT(1/a(Cmax-Cmin))$公式压缩索引空间，$Cmix-Cmin$ 首先把索引空间压缩为0～$Cmax-Cmin$ 个。倘若还大，取 $a>1$进一步压缩，但这可能使规则因 $F(Mi)=F(Mj)$而采用链接方式。

结束语 本文的近似评判方法主要辅助评选科技成果奖，评委可以对计算机直接计分提出意见，可以采纳或不采纳计算机对某一成果作出的建议，也可以输入某些数据干予计算机。这就增强了科技成果评奖的可操作性，提高了公正性，避免了一些人为因素。

本文所述的智能接口以子模块的形式与 MIS 并接。自然语言接口目前能满足用户提出的检索用语要求。当然，这种特定语言集合下的微机理解和处理相对于自然语言集来说只能算作探讨，但它对自然语言接口的实用提供了途径。

我们在488条原文分析规则集中进行了实验，附加数组容量为500，结果有链23个，均≤7，这些链是由 OR 条件，$F(Mi)=F(Mj)(i\neq j)$以及压缩空间引起，这与我们的分析是一致的，说明该算法行之有效。

通过实验，已证实本文所述算法具有实用性。这说明在特定环境中研究 MIS 智能化问题是有意义的。智能技术的应用是当今软件设计面临的一大课题，如果注意设计环境，就可以在待定场合中使用，为普及人工智能技术逐步总结出经验。

本文介绍的方法与常规的 MIS 相比，增加了软件设计难度，附加了一定的存储开销。因此，这种方式适合于较大的系统，如情报检索、图书管理等。而对仅千人以下的人事档案管理系统等，这未必是设计的最佳选择。

希望本文的算法和方法能视为 MIS 在常规检索方面发展的一个环节。或从某种意义上讲，本文的算法和方法可能为 MIS 在常规检索方面的发展提供启示。

参 考 文 献

1 Michacl RG, Nils JN. Logical Foundation of Artificial Intelligence. Morgan Kaufmken Publishers, Inc, 1987

2 Nguyend T, Vindrow B. Nearal networks for Self-Learning Control Systems. IEEE CSM, 1990, 10(3): 18～23

3 薛华成. 管理信息系统. 北京：清华大学出版社，1993, 6: 1～26

4 杨宪泽. MIS 的一个检索算法和纠错接口. 科技通报，1998, 14(2): 120～123

5 杨宪泽. 一种组织和检索文献的方法. 知识工程，1992(1): 46～59

6 杨宪泽. 直接映射式字符检索算法. 中文信息学报，1991, 5(3): 59～64