一个新的中文信息分类检索模型与系统实现

彭甫阳　何新贵
北京系统工程研究所
9702信箱19号，北京100101

摘要：本文首先提出一个基于知识的中文信息分类检索模型。该模型试图以概念为基础，以从文献中抽取的信息为证据，运用模糊推理技术，实现基于概念的中文信息分类与检索。其知识表示采用基于代数格和模糊数学的概念体系方法，其推理机制为证据驱动的模糊推理，可提供多种模糊算子完成合取、析取与传递。然后我们介绍一个基于该模型而实现的智能化中文信息分类与检索系统GD的总体结构。最后，简要介绍将该系统应用于一个具体领域所获得的一些试验结果及初步分析。

A New Model for Chinese Information Classification and Retrieval

Fu-Yang Peng and Xin-Gui He
Beijing Inst. of Systems Engineering(BISE)
P.O.Box 9702-19,Beijing 100101

Abstract: Presented in this paper is a new model for Chinese information classification and retrieval. The model is a knowledge-based one, which takes the information extracted from documents as evidences, uses knowledge base of conceptual hierarchy and fuzzy inference engine. Knowledge is represented in a so-called conceptual hierarchy the basis of which is algebraic lattice and fuzzy mathematics. Implementation issues and experimental results are also briefly discussed in this papaer.

1 引　言

信息技术革命使得越来越多的以自然语言形式记载的文献不再以书本为载体，而是大量地存在于计算机的各种媒体上。如何有效地组织、维护和利用这些文献已成了能否有效使用这些人类文明长期积累的丰富宝藏的关键。就汉语而言，在基本上解决了汉字的机器表示和汉字的输入输出问题以后，怎样在词、句乃至篇章等更高层次上实现对汉语的处理成了中文信息处理研究人员关注的重点。

计算机对自然语言的处理可以在许多不同的层面进行，从最简单的频数统计到复杂的

篇章理解形成了一个语言处理谱系。在这一谱系中的不同点上，我们所追求的目标是不同的，解决问题的方法是不同的，实现的难易程度和当前的成熟情况也是不同的。

本文的讨论定位在智能中文信息检索这一层面，即用基于知识的方法实现中文文献的自动分类和检索。

传统的信息检索一般是以主题词表为基础的，文献的内容以主题词表中的词组合来表征和标引，用户的检索要求也是以主题词表中的词构成的某种表达式（如布尔表达式）表达的。这类系统的查准率比较高，但查全率往往不理想，更严重的问题是，对标引者（人或自动机器）的要求高，并且主题词表由于其固有的特性不能反映变化了的情况。

全文信息检索与传统的信息检索正好相反，它并不需要主题词表，而是以文献中的词或词的组合来表征文献内容的。因此，对标引者要求不高，查全率一般比较高，但查准率取决于用户，这意为着增加用户的负担，是不能接受的。

本文讨论的GD系统是一个智能化的中文检索与分类系统。我们试图以概念为基础，用基于知识的方法实现中文文献的检索与分类。该系统结合了传统信息检索方法和全文信息检索方法的优点，但避免了两者的短处。与目前国内的全文检索系统比较，由于采用了基于概念体系的知识表示方法和模糊推理技术，本系统实现了全文概念（主题）检索，而其它全文检索系统只能实现全文符号检索。

2 GD系统的建模

2.1 GD系统模型

GD系统是一个基于知识的中文信息检索系统。我们试图以概念为基础，以从中文文献中抽取的信息为证据，运用模糊推理技术，实现基于概念的中文信息检索。

2.1.1 概念体系

概念体系是对概念的描述和组织，是信息检索中领域知识的表示。自然语言的概念之间存在各种各样复杂的关系，如同义、反义关系，上位、下位关系，组配、去配关系，具体、抽象关系等。概念之间的关联程度亦不相同，有些概念联系紧密，有些则不然。这就要求概念体系具有较强的表达能力和灵活性。

概念体系并不要求是对应的领域的完全描述。它只反映开发领域知识的用户或用户组的特定兴趣。用户可以构造自主的语义结构。同时，概念体系描述的知识并不要求一定是专家知识，当然，如果用户确实是专家，那么所开发的概念体系结构，随着时间的推移，最终可能成为该领域的专家描述。

概念体系的形式定义在[1]中给出。

2.1.2 GD系统模型

设$\sum = \{a_1, a_2, \ldots a_n\}$为汉字集合，$f_w: \sum^* \to \{0,1\}$判定一汉字串是否为一汉语词。$W=\{\alpha \mid \alpha \in \sum^*, \text{且} f_w(\alpha)=1\}$为汉语词的集合。又设要处理的文献集为$D=\{d_1, \ldots d_m\}$。要考察某一文献是否与某一概念相关联，我们可以按如下的系统方法进行。首先对文献d_i

分词，按某一系统词表（如果有的话）提取文献中的证据，然后用这些证据对概念体系（知识库）进行推理和求值，最后求得概念c与文献d_i的关联程度。由于概念与文献的关联是非精确的，因此，我们在求值的过程中应用的是模糊推理技术。

GD系统模型正是基于上述思想。下面是GD系统模型的定义。

定义1 GD系统的模型为一个十元组

$$<D, W, SL, KB, OP, V, Q, f_{seg}, f_{ev}, f_c>$$

其中，D为系统所处理的文献集合，W为汉语词集合，SL为系统词表集，KB为概念体系知识库，OP为模糊推理用选项集合，集合中的每一元素为四元组（and_op, or_op, detach_op, threshold），这里and_op为交型算子，or_op为并型算子，detach_op为蕴含算子，threshold为筛选阈值. V=[0, 1]. Q为由概念组成的集合。$f_{seg}: D\to P(W)$为分词函数，$f_{ev}: P(W)*SL\to P(W)$为证据提取函数。$f_c: P(W)*Q*KB*OP\to V$为概念求值函数。

下面我们对这一定义作进一步的说明和解释。

有两种类型的系统词表：非用词表和重要词表。证据提取函数的行为与词表的类型密切相关。若设文献D的词集为Wd，系统词表为sl，则

$$f_{ev}(Wd, sl)=\begin{cases} Wd-sl & \text{若sl为非用词表} \\ Wd\cap sl & \text{若sl为重要词表} \\ Wd & \text{若sl为空} \end{cases} \qquad (2.1)$$

概念c关于文献d的关联值为

$$\alpha_{cd}=f_c(f_{ev}(f_{seg}(d), sl), c, kb, op) \qquad (2.2)$$

式中$d\in D, sl\in SL, c\in Q$, $kb\in KB, op\in OP$.

在文献的检索和分类过程中，(2.2)式中的sl, kb和op一般已经固定. 因此，我们将(2.2)式简记为(2.3)式的形式.

$$\alpha_{cd}=f_c'(c, d) \qquad (2.3)$$

记D上的所有模糊集合为$\tilde{S}(D)$，C上的所有模糊集合为$\tilde{S}(C)$，关于GD系统中的文献检索和分类有如下定义。

定义2 概念检索$\tilde{f}_r$: $Q\to\tilde{S}(D)$定义为

$$\tilde{f}_r(q)=\{f_c'(q, d_1)/d_1, \ldots, f_c'(q, d_m)/d_m\}$$

定义3 文献分类$\tilde{f}_c$: $D\to\tilde{S}(C)$定义为

$$\tilde{f}_c(d)=\{f_c'(c_1, d)/c_1, \ldots, f_c'(c_n, d)/c_n\}$$

2.2 指标模型

对一般信息检索系统的性能评价，除费用指标，如响应时间和吞吐量之外，还有两个重要指标：查全率和查准率。查全率反映检索系统的漏检情况，而查准率则反映检索系统的误检情况。

在GD系统中，分类和检索的结果均为模糊集。假设D为被检索的文献集合，C为分类

体系中的类集合，专家认定概念s与D中文献关联的情况为模糊集$\tilde{P}_s^*$，系统给出的概念s的检索结果为模糊集$\tilde{P}_s$，专家认定文献d的分类结果为$\tilde{R}_d^*$，系统给出的文献d的分类结果为$\tilde{R}_d$，则系统分类和检索的性能可用两个模糊集的贴近程度表示，即

分类：$\| \tilde{R}_d - \tilde{R}_d^* \|$

检索：$\| \tilde{P}_s - \tilde{P}_s^* \|$

在模糊数学文献中，有各种计算贴近度的方法，如切比雪夫距离，均方差等。但这些方法有一个公共的缺点，计算的结果对信息检索来说无直观的解释。下面我们给出一种新的定义，使贴近度的计算与漏检率和误检率联系起来。

令$D_r=Supp\tilde{P}_s^*$，$D_\lambda=\lambda\triangle\tilde{P}_s=\{d_i \mid \mu(d_i)\geqslant\lambda\}$，分别表示$\tilde{P}_s^*$的支集和$\tilde{P}_s$的$\lambda$截集.显然有

命中数$=D_r\cap D_\lambda$

误检数$=\overline{D_r}\cap D_\lambda$

漏检数$=D_r\cap\overline{D_\lambda}$

系统的漏检率和误检率分别由以下两式给出：

$$h_r=K_m/K=|D_r\cap\overline{D_\lambda}|/|D_\lambda|$$

$$h_p=K_f/K=|\overline{D_r}\cap D_\lambda|/|D_\lambda|$$

检索误差定义为由漏检率和误检率组成的序偶。

$$\| \tilde{P}_s - \tilde{P}_s^* \| = (h_r, h_p)$$

由以上各式可知，漏检率和检索误差均与λ密切相关。现在我们再就λ的选择作进一步的讨论。

1) 若取$\lambda=\min\{\mu(d) \mid d\in D_r\}$，则有$h_r=0$，这时的$h_p$表示了在系统没有漏检的情况下的误检率，这实际上就是系统的查准率，它反映了系统排除非相关文献的能力。

2) 若取$\lambda=\max\{\mu(d) \mid d\in\overline{D_r}\}$，则有$h_p=0$。这时的$h_r$表示了在系统没有误检的情况下的漏检率。实际上这就是系统的查全率，它反映了系统选择相关文献的能力。

由于以上λ的取值均与专家认定的模糊集有关，在实际应用中难以获得。因此，在实际的系统中，控制输出结果的方法有二。一是给出某个指定阈值，其关联值超过该阈值的文献便作为结果输出；一是给出某个值m，将$\tilde{P}_s$按关联值的大小排序，取出对应前m个最大关联值文献作为结果输出。

类似地，我们可以定义关于分类的漏分率、误分率和分类误差。

3 GD系统总体结构

根据上一节给出的GD系统模型，我们可以设想GD系统逻辑结构应包括中文书面语分词子系统、文献预处理与证据生成子系统、概念体系知识库及推理子系统以及文献分类

与检索子系统。由于以往的中文信息检索系统多在字一级处理，在词一级水平进行的中文信息检索还处于探索阶段，远未达到成熟，见诸文献的成功例子极少，因此，我们在设计GD系统的结构时将不囿于现有的中文信息检索技术，而是以一种统一的观点来设计GD系统的构架，在这一构架下既尽量利用已有中文信息检索技术的成果，又深入分析当前技术存在的问题，提出我们自己的解决办法和实现途征。

3.1 GD系统的设计目标

GD系统的设计目标是综合利用数据库技术、人工智能技术、模糊技术及中文信息处理等最新技术，吸收传统的主题检索和全文检索的优点，实现词一级的基于概念的中文信息检索。系统应具有足够的灵活性，能选择不同的工作参数，能配置成不同的工作模式。系统应具有友好的用户界面，并且能对不同类型的用户展示不同的界面。系统应具有集成多种功能和工具的能力，除核心功能外，还应提供诸如列表、浏览、编辑、索引重建等功能。

3.2 GD系统的结构

GD系统可以分成四层(如图3.1)。第一层为支撑层，包含数据库支撑子系统 DBMS/GD和中文用户界面支撑子系统CWIN/GD。数据库支撑子系统提供系统所需的模式定义、数据操作、索引管理等功能。中文用户界面支撑子系统提供窗口、菜单、消息框、控制盒等用户界面对象。

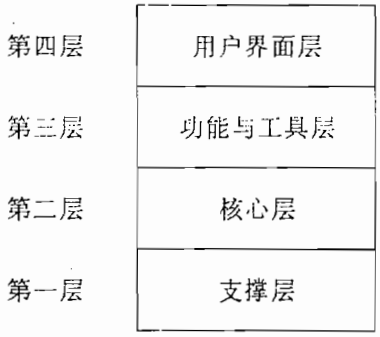

图3.1 GD系统的结构

第二层为核心层。核心层实现GD系统的核心功能。包括文献的分词，文献的预处理与证据生成，概念体系知识库的管理，以及基于模糊推理的概念求值等子系统。文献分词实现将字串形式的文本到词串的转换。文献的预处理与证据生成子系统对分词后的文献进行处理，识别出词及其位置特征。如标题、段号、句号和词号等，根据系统词表的类型与内容，决定是否将其作为证据插入证据倒排库中。概念体系知识库管理子系统主要负责知识库内外形式的转换、知识库的装载及知识库内存映象的管理。概念求值子系统根据知识库中已有的知识和证据库所提供的证据进行模糊推理和计算。系统提供多种模糊算子组合，

提供控制结果输出的手段，提供基于记忆的优化技术。

第三层为功能与工具层。本层实现文献的登录、文献的分类与检索等功能，提供列表、浏览器、编辑器、用户管理、选项管理、索引重建、联机帮助等工具。本层是GD系统功能的汇集。

第四层为用户界面层。该层实现以友好的方式提供用户操作与使用GD系统功能的中文用户界面。

4 GD系统应用于提案处理

4.1 知识获取

GD系统可用于提案处理中与分类和检索有关的问题，如查找某一主题的提案，对提案按主题词表分类，确定提案的承办单位和会办单位等。要实现这些功能，就需要获取与提案处理有关的知识，这些知识大体上可分为以下几类：一、语言中的语义知识；二、分类知识；三、机构及其职能的知识。

语义知识给出各种实义词之间的各种关系，主要包括同义反义关系，上位下位关系，组配去配关系等关系。

分类知识涉及到提案处理中与分类有关的知识，如类的划分，类的层次等。

机构与职能知识描述可能的提案承办单位的职能，以及与提案处理的关系。这些知识用来确定提案的承办单位和会办单位。

这些知识经编码形成系统知识库。目前系统中分类规则为175条，领域规则为311条。

4.2 实验结果分析

我们对全国政协提供的１００多篇提案进行了分类和检索实验，获得了令人满意的结果。这批提案涉及政治法律、人事福利、计划统计、财贸金融、文化宣传、教育事业、科学技术、轻重工业、医药卫体、交通邮电、城乡建设、农林牧业、统战综合等１３个大类，我们用GD系统的文献分类功能将它们归类到政协提供的分类表中，其中模糊推理选项取系统缺省值。实验结果表明，分类完全正确和基本正确的文献占总文献数的９１％，分类错误（包括错分和漏分）的文献占９％。这部分分类错误的原因主要有以下几个方面：

１、文献本身的原因，如提案正文过短，提案叙述方式使主题发散，提案用词与主题偏移太多等。我们的系统是一个以证据为基础的推理系统，证据获取的正确与否直接影响分类和检索的结果，而证据基本上是按概念词在文献中出现的位置和出现频数的某种加权运算，如果Zipf定律的条件不能满足，则分类和检索的结果就可能出现偏差。

２、知识库不完备。知识库中的知识提供了各概念之间的相互关系以及概念与证据、概念与其它信息源的关系。知识越丰富，则推理的深度和广度就能增加，从而就能提高分类结果的正确率。但目前的知识库规模还较小，知识还很不完备。

３、分类体系（分类表）存在的问题。我们得到的分类表主要存在两个方面的问题。一是个别分类主题词专指性较差，如“城乡建设”大类下的“建设”和“科学技术”大类

下的“信息”主题词等；二是个别大类下的小类似乎缺一些主题词，如“农林牧业”大类下似应包括“农业”这样的主题词。

4、分词词典不完全造成一些概念词的切开，如少数民族的名字切散后就可能使一些与少数民族有关的提案难以分到“民族”这一主题类下。虽然这可以在知识库中通过加入组配规则来弥补，但这将大大增加知识库中的规则数目。

5、模糊推理选项的影响。算子和阈值均可能对推理结果造成影响。实验表明，对大多数较规范的文献，推理选项对结果在质上影响不很明显，虽然量上有些变化。但对上面提到的某些非“规范”文献，算子的选择可能在质上影响结果。如选择突出主因素的算子就不如多因素综合的算子。

5 结束语

本文描述了一个智能化中文信息检索系统。该系统是在DOS平台上用C/C++语言开发的。为使系统进一步实用化，还应该考虑以下几方面的工作。一、对系统进一步完善，如工具层的编辑器和浏览器，功能层的检索功能都应该改进和加强；二、本系统的证据在很大程度上仍然是统计意义上的，对文献的预处理是表层的，基本没有进行深层的结构分析。如果能引进自然语言的一些处理技术，肯定会对证据的质量及分类检索的结果带来积极的影响；三、将该系统移植到WINDOWS平台，以改善用户界面；四、考虑在UNIX平台和TCP/IP网络环境下实现基于客户--服务器的分布式智能信息检索。五、实现多媒体信息的一体化管理与检索。

参 考 文 献

[1] Peng, Fu Yang and He, Xin Gui, Conceptual Hierarchy: Formalism, Implementation and Applications, Proc. 1992 Int. Conf. on Chinese Information Processing, vol. 1, 1992

[2] Peng, Fu Yang, et al., A Knowledge-Based Document Retrieval and Classification System, Proc. Pacific-Asian Conference on Expert Systems, 1995

[3] Peng, Fu Yang and He, Xin Gui, Text analysis in Information Retrieval Environment, Proc. Int. Conf. for Young Computer Scientists, 1995

[4] 彭甫阳，冯京，何新贵，一个提案分类检索专家系统的设计，首届全国计算语言学联合学术会议论文，1991

[5] 彭甫阳等，概念体系研究，第四届全国青年计算机工作者会议论文集，1992.

[6] 彭甫阳，何新贵，文本分析与信息检索，第二届全国计算语言学联合学术会议论文，1993

[7] 彭甫阳，何新贵，一个基于概念体系模型的全文文献检索分类系统的设计与实现，第3届全国人工智能联合会议论文集，1994

[8] 彭甫阳等，GD--一个智能化中文信息检索系统，国防系统分析与软件，95年第2期