蒙文单词全文检索系统 MWTSS 的设计

宋建斌[1]，敖其尔[2]，马丽[3]

（内蒙古大学 计算机学院，内蒙古 呼和浩特 010021）

摘要：本文首先分析了该课题的背景知识和国内外研究现状，在分析蒙文单词全文检索系统 MWTSS（Mongolic Word Text Search System）系统的特点和功能之后，重点给出了 MWTSS 的系统设计，介绍了 MWTSS 的整体结构，分析了 MWTSS 全文数据库，建立索引子系统和关键词检索子系统的基本结构。最后，对本课题过去的和将来要做的工作做了总结。

关键字：全文检索；蒙文单词；索引表；数据库

中图分类号码：H085.6　　**文献标示码：**A

全文检索，是指以文本信息作为检索对象，建立全文检索数据库，除具有布尔逻辑检索功能以外，还具有文本检索功能。并允许用户以单词或词组检索，直接获得原文中的有关章节和段句。在信息检索领域，全文检索一直是一个比较复杂的问题。与普通数据库检索所涉及的结构化数据查询不同，全文检索不仅要查询结构化数据，而且还要查询非结构化数据。比其标引检索，全文检索提供了全新的、强大的检索功能，能够方便多角度、多侧面地综合利用信息资源。

我国计算机全文检索起步于 80 年代初，并在计算机编制主题词表、汉语自动分词和标引、数据库建造、图书馆情报、全文检索理论等领域取得了很大进展。国内外相继出现了大量的中英文全文检索软件系统，英文和中文的全文检索技术得到了长足的发展。如北大方正的 MIRS、上海交通大学王永成教授主持的“ 法律条文全文数据库 ”、陕西省中医研究院研制成功的“ 中医经典古籍全文数据库 ”、深圳大学建立的“ 古典名著《红楼梦》全文数据库 ” 等。全文检索技术愈来愈成为信息系统中不可缺少的关键功能和必要手段。

随着计算机技术在蒙文应用领域的不断拓展，由计算机管理的蒙文文档资料的数目也飞速发展。蒙古语言文字是我国少数民族蒙古族使用的语言文字，是内蒙古自治区的官方语言，在国际上也是很有影响的语言文字。在内蒙古自治区有大量的、完整的、全面的，十分珍贵的蒙文信息资源。国家在内蒙古自治区建立了“ 国家教委民族学科蒙古学文献信息中心 ”，具有图书资料 10 万余册。在社会科学方面，有蒙古语言文学、蒙古文学、蒙古古代、近现代史研究，而且与国际上许多国家（如：日本、美国、加拿大、德国、俄罗斯、蒙古国等）和地区有着十分广泛和密切的有关蒙古学的科研交流。为了能有效的存储和管理如此庞大的蒙文信息资源，对蒙文全文检索系统的需求日益加剧。

在内蒙古自治区，已经有 3 万多用户接入 Internet，目前正以越来越快的速度发展。但使用的网络资源多为中文信息和英文信息，而本地区的主体民族语言——蒙文信息在网上几乎是没有，原因是目前还没有支持蒙文信息上网的基本技术。1999 年国家正式启动“ 政府上网 ” 工程，蒙文作为自治区官方文字，在政府办公过程中必不可少，蒙文信息若不能上网，势必严重阻碍该工程的实施和政府办公。蒙文全文检索技术是支持蒙文上网的基本技术之一，对它的研究就显得尤为重要。

然而，当前对蒙文全文检索技术的研究还处在起步阶段，基于此，我们研制和初步实现了蒙文单词全文检索系统 MWTSS，以满足日益增长的需求，填补该研究领域的空白。以此为蒙文上网工程提供必要的技术支持。

一、蒙文单词全文检索系统 MWTSS 概述

将文本文献变为有意义的单词序列，这在信息处理和检索技术中是由自动分词技术来实现的。蒙古语言的书写形式和英语、汉语都不相同，它是以词为单位，按列向下书写的，词和词之间以标点符号或空格分隔开。但在计算机的内部表示上和英文相似，蒙文中的列被转换为行，存贮在计算机中，英文的词和词之间使用标点符号或空格分隔开。所以，蒙文原文档的分词技术和英文文档的分词技术相类似。蒙文的词法结构较为复杂，一个字母在词首、词中、词尾有三种不同的写法，同音不同形、同形不同音的现象很普遍。蒙文词是由词根、词干、附加成分、词缀组成，蒙文静词类有数、格、领属的变化，动词的组词和形态变化更为复杂，因此在使用蒙文单词进行检索词的一个关键性问题是寻找该词的词根。初步的设想是建立一个蒙文规则知识词库，当用户输入关键词进行检索时，首先根据检索规则知识库对关键词进行截词、派生、变异，找到该关键词的词根，再用该词根进行模糊查询，以提高查全率。

我们的研究目标是综合多种检索模型，研究和建立一个适合蒙文的、查全率和查准率都较高的检索模型。和中、英文的全文检索系统相同，我们设计的蒙文单词全文检索系统 MWTSS 的最基本的功能，是可以根据用户输入的蒙文关键词，检索出用户所需的蒙文文档信息，并把结果反馈给用户。查准率表示检索的准确程度。查全率表示检索的结果是否全面。为提高检索结果的查准率和查全率，在全文检索系统的设计过程中，采取了一些相应的技术，如采用后控词表来完成对同义词和近义词的检索控制，在检索过程中根据检索规则知识库对关键词进行截词、派生、变异等。图 1 是本系统的简要功能流程图。

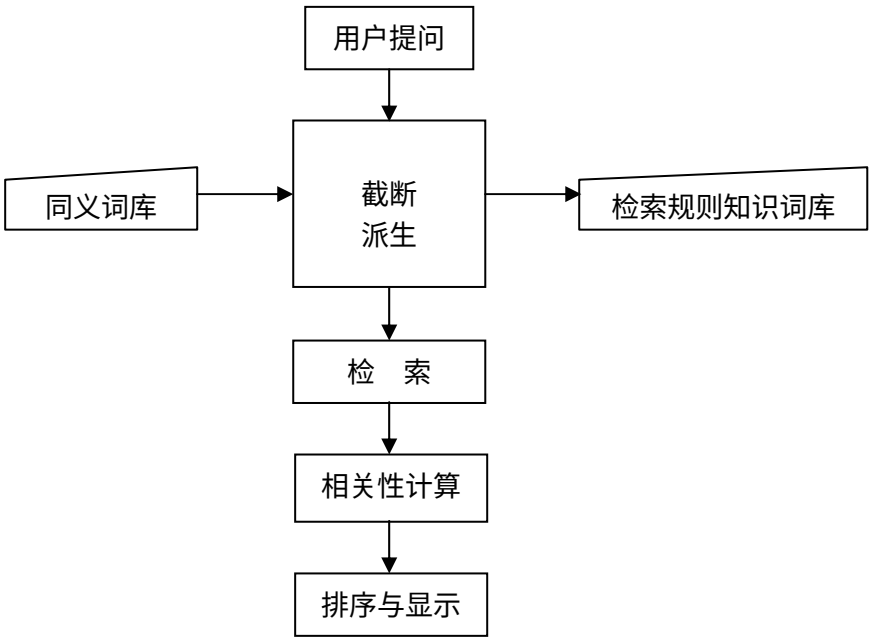

图 1：MWTSS 系统简要功能流程图

本系统的主要特色和创新之处是：

（一）结合蒙文特点，研究蒙文全文检索技术；综合多种检索模型，研究和建立一个适合蒙文查全率和查准率都较高的二级倒排表检索模型。

（二）用同义词链表结构取代后控词表，提高了检索效率，并使系统具备了初步的学习功能，提高了智能化程度。

（三）在关键词检索模块中使用模糊查询技术，以此提高检索的查全率。

（四）为建立蒙文信息服务站提供必要的技术保障。

二、蒙文单词全文检索系统 MWTSS 设计

本节先介绍了蒙文单词全文检索系统 MWTSS 的整体结构和 MWTSS 全文数据库的基本结构，然后对蒙文单词全文检索系统 MWTSS 各子系统的设计进行详述。

（一）蒙文单词全文检索系统 MWTSS 的整体结构

从功能的角度看，蒙文单词全文检索系统 MWTSS 可分成两个子系统，即生成索引子系统和关键词检索子系统。生成索引子系统和关键词检索子系统分别完成预处理文档生成索引库和根据用户输入的关键词进行检索的功能。其中，生成索引子系统可分为三个功能模块，分别是文本格式转换和非蒙文单词的过滤模块、预处理蒙文文档生成索引库模块和停用词库的处理模块。关键词检索子系统可分为四个功能模块，分别是生成蒙文词根模块、生成近义词链表模块、关键词检索模块和检索结果反馈模块。图 2 是蒙文单词全文检索系统 MWTSS 的层次结构图。

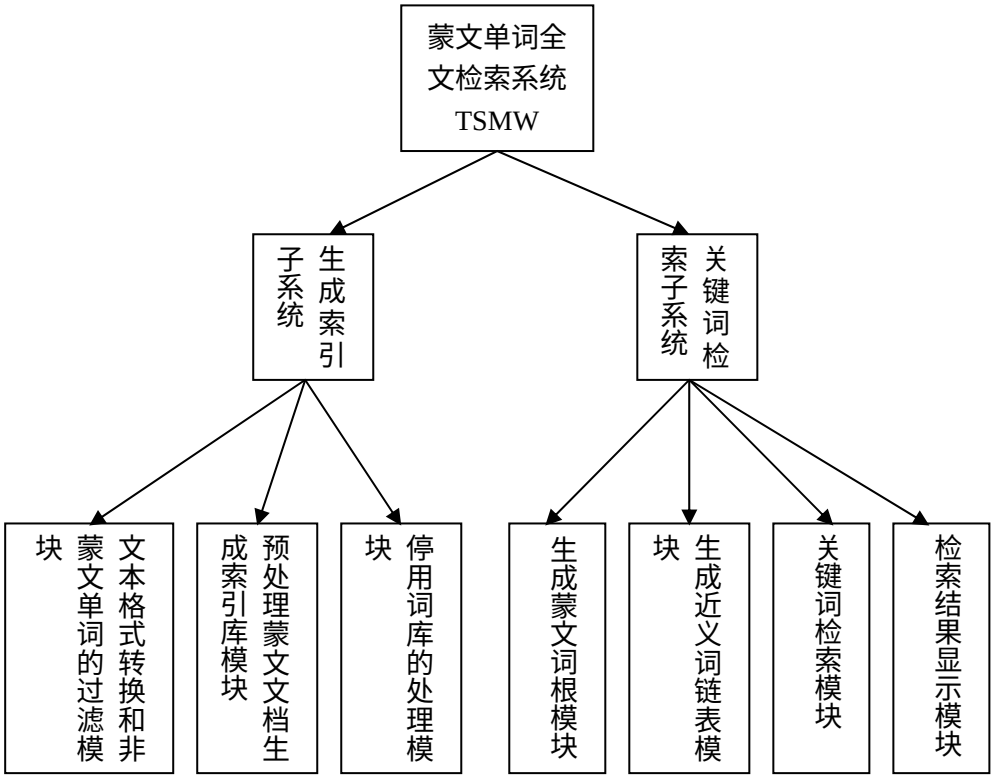

图 2：MWTSS 系统的层次结构图

（二） 全文数据库的基本结构

任何一个全文检索系统都应维护一个全文数据库，用来存储正文文献，文献索引等内容。蒙文单词全文检索系统 MWTSS 的全文数据库由三个表组成，分别是：

正文表：用来存储蒙文文档。其关系模式为：正文表（文献号，正文）。其中“ 文献号 ”是每个文献的唯一标识，由系统自动编号。“ 正文 ”字段用来存储该文献的全部或部分内容。其中“ 文献号 ”字段是正文表的主键，正文表用它和索引表相关联。

索引表：用来存储对正文表中各文献的索引信息。由于蒙文文献类似于西文文献，各单词之间由空格和标点符号分开，很容易切分，所以索引表采用基于词索引法的二级倒排表结构。

一级索引表的关系模式为：一级索引表（序号，关键词，指针）。其中“ 序号 ”是每个关键词的唯一标识，由系统自动编号。在一级索引表中，“ 序号 ”和“ 关键词 ”字段的内容是唯一的，即在表中不可能出现两个相等的序号值，也不可能出现两个相同的关键词。“ 指针 ”字段用来形成同义词循

环链表。

二级索引表的关系模式为：二级索引表（序号，文章标题，文章作者，文献号，权值)。其中的“ 序号 ”字段和一级索引表中的“ 序号 ”字段一起把一级索引表和二级索引表联系起来 ;“ 文章标题 ”和“ 文章作者 ”字段分别存储关键词所在文献的标题和作者 ;“ 文献号 ”字段对应于正文表中的“ 文献号 ”字段，表示关键词所在文献在正文表中的相对地址，用于显示文献正文 ;“ 权值 ”字段用来表示该关键词在本文章中出现的频率。

停用词表：用来存储那些蒙文中没有实际意义的词汇。其关系模式为：停用词表（停用词，标记)。其中“ 标记 ”字段用来表示是否用该停用词更新过索引表。

在索引表的构造过程中和利用索引表进行检索时采用了两项关键技术：

1、 词加权技术：在对文档进行预处理创建索引的过程中，采用词加权技术来表示词汇在文档中出现的频率，在文档中出现频率越高的词，其在索引表中的权值越大，在检索时，检索结果按权值从大到小的顺序输出。

2、用同义词链表取代后控词表：在本系统中，用在一级索引表中构造同义词循环链表的方法取代普通全文检索系统中的后控词表，即在一级索引表中设置一个“ 指针 ”字段，该字段利用“ 指针 ”字段可以在一级索引表中形成循环链表，把所有的同义词都链接在一起。在检索时，通过检索一级索引表得到检索词的所有同义词，再利用检索词和其所有同义词的序号对二级索引表进行检索，即得到所需的检索结果。这种方法和普通的设置后控词表的方法相比，减少了对数据库中表的访问次数，既节省了系统的空间开销又提高了检索效率，是对检索方法的一种有效的简化。

（三） 建立索引子系统

1、 功能描述

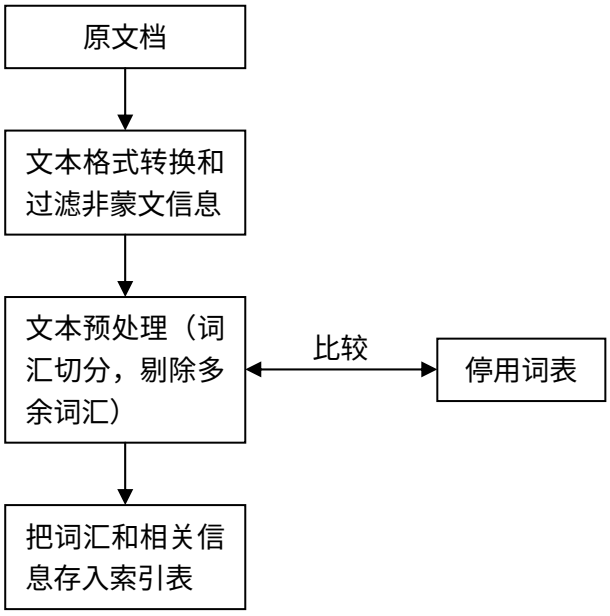

图 3：MWTSS 系统索引创建流程图

全文检索的核心技术就是将原文档中所有的基本元素的出现信息提取出来记录到索引库（表）中，为该文档创建索引。蒙文单词全文检索系统 MWTSS 的建立索引子系统所完成的功能就是对蒙文原文档进行预处理，把其中的基本元素（蒙文单词）的出现信息提取出来记录到索引表中，以供用户检索时使用。索引创建流程如图 3 所示。建立索引子系统从功能上又可分为三个模块：文本格式转换和非蒙文单词的过滤模块；预处理蒙文文档生成索引库模块；停用词表的处理模块。

2、 文本格式转换和非蒙文单词的过滤模块

本模块要完成的功能是：把 WORD 文档，WPS 文档等各种不同格式的蒙文文档转换成 Text 纯文本格式的文档，然后再把纯 Text 文档中的非蒙文信息过滤掉。

本模块主要是为进一步处理蒙文文档（即对蒙文文档进行分词处理）作前期的准备，因为待处理的蒙文原文档可能属于不同的类型，如 WORD 文档类型或 WPS 文档类型。而这些类型的文档中都存在大量的非正文信息——称之为标记，这些标记用来标识显示文档的各种属性。每种类型的文档都有自己独特的标记系统，如不经处理就会对这些源文档进行切分词，则得到的结果必然是错误的。这就要求我们在对不同类型的文档进行切分词处理之前，先要统一格式，即把这些文档中的标记删除，转变成不含标记的纯文本格式的 Text 文档。同理，还必须把原文档中的非蒙文信息也剔除掉。

3、 预处理蒙文文档生成索引库模块

本模块是建立索引子系统的核心部分，其功能有三：

（1）读取纯 Text 文本格式的蒙文文档，并按照各种标点符号和空格进行分词处理。

（2）注意处理分出的每一个蒙文单词。判断停用词表中是否存在该词，如存在，则忽略该词；如不存在，则继续。按单词在文章中出现的位置和频率对每一个单词进行加权（单词出现在标题中的权值大于出现在正文中的权值；单词在文章中出现的频率越高，权值越大）。

（3）创建索引。如蒙文单词已在一级索引表中，则从一级索引表中读取该词所对应的序号，并连同文章题目，文章作者，文献号和该词的权值一起存入二级索引表中；如该单词不在一级索引表中，序号值为原最大序号值加 1，指针值和序号值相等。然后，再把该序号连同文章题目、文章作者、文献号和该词的权值一起存入二级索引表中。

4、 停用词表的处理模块

停用词表在全文检索中具有极其重要的意义和使用价值，利用它可以有效的减少系统数据冗余，提高检索效率。本模块的功能是添加和维护停用词表，并使用它对索引表进行更新。主要有以下两个部分：

（1）添加停用词表。这部分功能只有系统管理员才能使用，可以在系统实现的初期进行添加，也可以在系统的使用过程中进行添加。如某停用词已在表中，则不添加；否则添加，并置“ 标志 ”字段的值为 0，表示还没有用该停用词更新索引表。

（2）定期用停用词表对索引表进行更新。把索引表中的那些无实际意义的无须作为关键词的蒙文单词删除以减少数据冗余。把停用词库中已更新过索引表的停用词所对应的“ 标志 ”字段的值置为 1，表示已用该词更新过索引表，避免在下一次更新时进行重复的操作。由于在更新过程中，会引起一级索引表中“ 序号 ”字段的值的改变，所以需重建一级索引表和二级索引表，特别要注意的是要重建一级索引表中的近义词循环链表。

（四） 关键词检索子系统

1、 功能描述

除了创建索引技术，在全文检索中的另一项关键技术是关键词检索技术，即根据用户输入的关键词，采用适当的策略对全文数据库进行检索得到用户想要的信息。

蒙文单词全文检索系统 MWTSS 中的关键词检索子系统所完成的功能就是根据用户输入的蒙文关键词，采用适当的策略在全文数据库中进行检索，并把检索到的信息反馈给用户。其基本流程如图 4 所示。

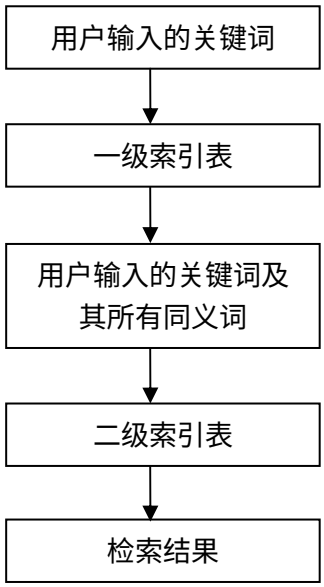

图 4：MWTSS 系统关键词检索流程图

关键词检索子系统从功能上又可分为四个模块，即生成蒙文词根模块、生成近义词链表模块、关键词检索模块和检索结果反馈模块。

2、 生成蒙文词根模块

蒙文单词的构词法和英文的类似，即可以在一个词根后面加上不同的后缀来形成词性、时态和格的变化。针对蒙文的这种特点，我们采用了模糊查询技术来提高全文检索系统的查全率。以下就模糊查询技术作一个简要的介绍。

模糊查询技术：数据库应用系统提供了一种模糊查询技术，即可以在带查询的关键词字符串中的任何位置加入符号%，并使用比较运算符“ like ”替换查询语句中的比较运算符“ = ”来完成模糊查询。其中%符号代表由任意字符组成的字符串。如我们可以在字符串 abcd 后面加上符号%形成新的字符串 abcd%，并使用如下语句对字符串 abcd%在数据库表 Table 中进行检索操作：

```
SELECT * FROM Table WHERE 关键词 like ’abcd%’
```

则检索的结果是 Table 表中“ 关键词 ”字段的值等于前缀为 abcd 的字符串的所有记录。通过使用模糊查询技术，在进行全文检索时，我们可以在蒙文检索词的词根后面加上符号%，就可以对由该词根派生出来的所有蒙文单词进行检索，以提高检索系统的查全率。这就要求我们首先要找到蒙文检索词的词根。本模块的功能就是利用适当的方法找出用户输入的蒙文检索词的词根。

3、生成近义词链表模块

本模块带有初步的学习功能，是对全文检索技术的一个创新。基本功能是在用户对关键词进行检索时，提示用户同时输入该关键词的同义词或近义词。如用户输入了该关键词的同义词或近义词，则在一级索引表中创建近义词循环链表。如该链表已经存在，则进行必要的插入操作，即把用户输入的在同义词链表中上不存在的同义词插入链表。通过不断的使用，系统对同义词或近义词的检索能力也会不断的提高，检索的查全率就会逐渐提高。这种学习功能使系统具有了一定的思维、记忆和联想能力，提高了系统的智能化程度。但在实际应用中也存在一些问题，例如，如果同义词被用户错输特别是乱输，则会在索引表中造成混乱。不过，我们也可以提供一些必要的防范措施。例如，设置密码，只对可靠的用户提供该项功能。

4、关键词检索模块

查询技术是全文检索系统的核心技术之一，直接影响系统的执行效率和检索的查准率和查全率。所以本模块是关键词检索子系统的核心，基本功能是针对用户输入的关键词，对索引表进行检索以得到用户想要的信息。主要有以下几步完成。

（1）根据用户输入的蒙文关键词的词根，遍历其在一级索引表中的同义词链表，得到该词及其所有的同义词。

（2）对上一步中得到的所有词，利用模糊查询技术，在一级索引表中进行模糊检索以涵盖由该词根派生的所有蒙文词汇。

（3）用上一步中检索出的全部的蒙文单词所对应的序号，在二级索引表中再次进行检索以得到最后的检索结果。

5、检索结果反馈模块

本模块的主要功能就是把检索的结果清晰、完整的反馈给用户。其实现涉及到蒙文的输出技术，不在讨论范围内。

参考文献

[1] 姚天顺.自然语言理解[M].北京:清华大学出版社，1998.

[2] 清格尔泰.现代蒙古语语法[M].呼和浩特:内蒙古人民出版社，1998.

[3] 确精扎布.蒙古文编码[M].呼和浩特:内蒙古大学出版社，2000.

[4] 董春晓.万维网上的全文检索技术及其发展[J].情报理论与实践，2000,(1). 53-54.

[5] 曹元大,等.全文检索字索引技术的研究与实现[J].计算机工程，2002,(6).260-262.

[6] 马迎春.全文检索系统概述[J].情报科学，2000,(12).1132-1135.

Design of mongolic word text search system

SONG Jian-bin[1],Ochir[2],MA Li[3]

(College Of Computer Science Inner Mongolia University, Hohhot 010021)

Abstract: This paper first analyses the background and present conditions of studying and developing trends of the Editor for the Mongolian Scrip. When analyses characteristic and function of the Text Search of Mongolic Word, this paper emphatically introduces the design of MWTSS（Mongolic Word Text Search System）system, introduces the whole structure of MWTSS, analyses text database of MWTSS, then introduces the structure of the building index search sub-systems and the keyword search sub-systems. Finally, this paper gives a conclusion for past and future work.

Key Words: text search; Mongolic word; index table; database

收稿日期:2004-5-13;

基金项目：教育部人文社会科学研究重大项目“ 蒙古文信息处理平台（MIPP）的研究 ”(项目编号：02JAZJD850003)；国家自然科学基金项目“ 基于 WEB 的蒙文图书信息管理系统 ”(项目编号：60163003) .

作者简介：1，宋建斌（1977-），男，汉族，内蒙古呼和浩特市人。内蒙古大学计算机学院研究生，主要研究蒙文信息处理和计算语言学；

2，敖其尔（1941-），男，蒙古族，内蒙古兴安盟人。内蒙古大学计算机学院教授，内蒙古大学蒙古学研究中心兼职教授，硕士生导师，主要研究蒙文信息处理和计算语言学；

3，马丽（1978-），女，汉族，内蒙古鄂尔多斯市人。内蒙古大学计算机学院研究生，主要研究信息处理和数据库技术。