INNER MONGOLIA NORMAL UNIVERSITY

本科毕业论文（设计）

蒙古文网络数据爬取及语料库管理系统

Mongolian network data crawling and corpus management system

学 院 名 称：	计算机科学技术学院
专 业、年 级：	计算机科学与技术（师范蒙班）
作 者 姓 名：	娜何雅
作 者 学 号：	20172104256
指 导 教 师：	王斯日古楞

2021 年 4 月 27 日

毕业论文原创性声明

本人所提交的毕业论文“蒙古文网络数据爬取及语料库管理系统”，是在指导教师王斯日古楞老师的指导下，独立进行研究工作所取得的原创性成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中标明。

本声明的法律后果由本人承担。

论文作者（签名）：　　指导教师确认（签名）：

年　月　日　　年　月　日

目录

1 研究背景 1
2 开发工具简介 1
3 数据爬取和数据预处理 2
3.1 文本数据爬取 2
3.2 文本预处理及存储 3
3.2.1 文本数据预处理 3
3.2.2 数据库设计 3
4 蒙古文语料库管理系统 4
4.1 系统设计 4
4.2 系统功能实现 5
4.2.1 用户管理模块 5
4.2.2 语料库管理模块 7
5 总结 9
参考文献 9

蒙古文网络数据爬取及语料库管理系统

计算机科学技术学院　2017 级师范蒙班　娜何雅　20172104256

指导教师　王斯日古楞教授

摘要　本文通过网络爬虫技术获取了蒙古文文本数据,并对爬取的文本内容进行了预处理工作,其中包括去除无用的符号、对文本进行分句,将分句后的文本存入数据库中。然后,设计和实现了蒙古文语料库管理系统，系统主要有添加，修改，删除，查询统计等功能。本文首先介绍了文本数据爬取和预处理方法，然后给出了语料库管理系统的设计和实现。

关键字　蒙古文数据爬取;数据处理;语料库管理

1 研究背景

随着互联网的越发普及，网络爬虫技术作为数据采集的核心部分，成为了我们获得数据资源的主要方法。我们可以通过网络爬虫来获取大量的语言资源，这可以解决自然语言信息处理的资源建设问题。语言资源的缺少，很难将自然语言处理研究发展到真正的大规模和实用程度,所以语言资源在自然语言处理中起着重要的作用。因此，对蒙古文网站进行文本数据爬取也同样可以进一步扩展蒙古文自然语言处理方面的研究，提高相关工作者的工作效率。

随着社会的逐渐进步，互联网信息逐渐多元化，丰富化，蒙古文网站也逐渐变得复杂多样了起来。虽然网站的种类日益增加，但是目前蒙古文语料库的发展相对滞后，规模也有限，不能满足自然语言处理中的需要。本文通过网络爬虫技术获取了蒙古文文本数据。进而形成了一个蒙古文语料库管理系统，可以通过查询统计、添加、修改、删除等方式进行管理。

2 开发工具简介

本文的主要工作有三部分，第一部分是蒙古文文本爬取，第二部分是文本预处理及存储，最后一部分是蒙古文语料库管理系统。主要使用 Python 语言 3.9 版编译器，并在 Pycharm 软件平台进行一系列工作。

本文数据库设计采用的是 MySQL 数据库，它具备的优点是容量大，支持各种开发语言，如 PHP、Java、Python 等语言，本文编程语言使用的是 Python 语言。使用 Navicat Premium 软件连接到 MySQL 数据库进行蒙古文文本数据存储。

操作系统：Windows 操作系统

文本数据爬取及预处理：Python 语言 3.9 版本编译器及 Pycharm 软件

文本存储：Mysql 数据库及 Navicat 连接器

语料库管理：Pycharm 软件 Flask 库及 Mysql 数据库

3 数据爬取和数据预处理

本文首先进行了蒙古文网站的文本数据爬取工作，之后对爬取下来的文本数据进行文本预处理。预处理工作包括去除指定无用的符号，对蒙古文文本进行分句处理等操作。

3.1 文本数据爬取

本文爬取的内容是成吉思汗网（qinggis.net）的“小说”部分，共 30 页，1200 篇文章。文本数据爬取流程如下：从目标页面里面选择相应的 URL 链接，把选中的链接当作开始的 URL 链接。将这些链接存到新建立的等待抓取的 URL 链接队列。然后，顺序读取等待抓取队列里的网页链接，并判断这些链接是否已经被抓取完毕。如果抓取完毕则将这些链接存到已经抓取队列中，如未抓取则将链接存到未抓取队列中。然后对网页内容进行下载存储。以此类推，重复上述操作过程，直到抓取完等待抓取队列的链接，结束操作。爬取蒙古文文本数据详细步骤为如图 1 所示。

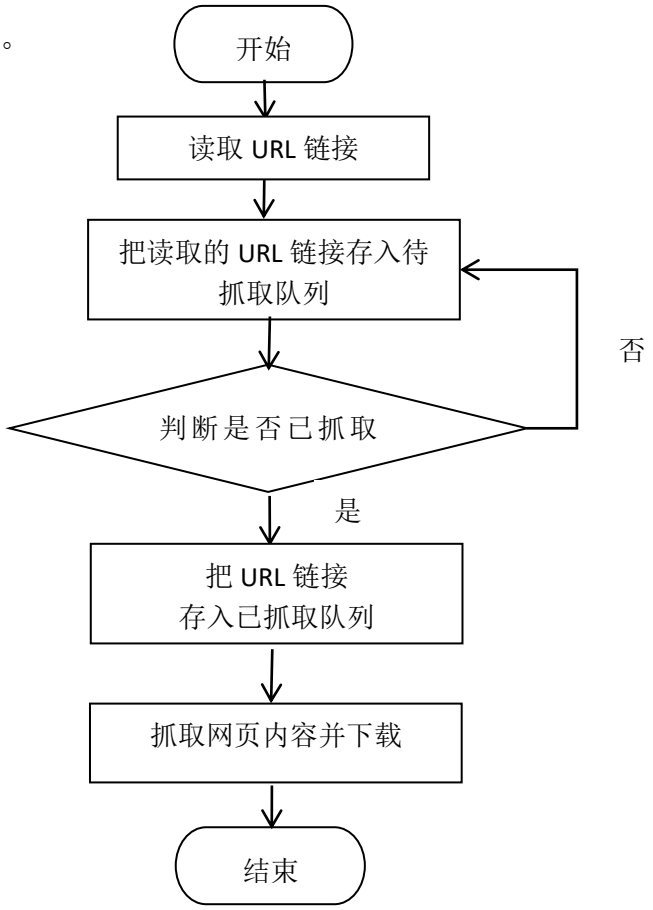

图 1 文本数据爬取流程图

本文在存储蒙古文文本的时候采用了 python 语言自带的 docx 库，并且对爬取的文本数据指定了字体“Menksoft”，对其进行存储。将爬取的每一篇文章都存储到一个文档里，共下载了 1200 个文档。

3.2 文本预处理及存储

文本数据进行预处理工作包括去除无用空格，去除无用换行符，去除无用英文符号，对其进行句子切分。将切分好的句子一条条存入到 MySQL 数据库的 test 数据库中。

3.2.1 文本数据预处理

本文主要用到了 python 语言的 replace 函数，此方法是将字符串中的 old 旧字符串，替换成 new 新字符串。如用 str.replace(‘old’,’ new’)，构成 str.replace(‘ ’, ‘’)语句，实现将空格去掉。其中，把之前存在 word 文档里面的文本数据转换成了 txt 文档，并将全部文档进行了合并处理，这样方便做预处理工作。

最后进行句子切分操作，先读取爬取下来的文本文档，再对文本数据顺序遍历，在此过程当中若是遍历到‘᠃’，‘!’，‘？’符号，则表示一个句子的结束。所以每当遍历到以上三个标点符号中的任意一个便换行输出，直到遍历完所有文本数据，将切分好的句子写入一个新文档里。以此类推，最后把切分好的句子存储到数据库中，形成了十万行语句。

3.2.2 数据库设计

本文选择了 Navicat Premium 软件与 MySQL 数据库进行了连接。将切分好的句子存入到 MySQL 数据库的 test 数据库中。在语料库管理平台中，我们需要用到两张数据表，一个是存放蒙古文文本数据的 testtxt 表，另一个是存放用户信息的 user 表。下面将具体的介绍每一个数据表的结构设计以及包含的数据内容。

（1）文本数据表

此 testtxt 表用来存放预处理分句后的文本数据，约有十万多行文本数据。表中分别有 id 和 process 两个字段名，id 为主键，属性自动递增，表示存入数据库的句子编号，process 表示每条句子相应的内容。如表 1 所示。

表 1 文本数据表

字段名	类型	长度	小数	不是 null	主键	内容
id	int	11	0	是	是	句子序号
process	varchar	1024	0	否	否	句子内容

（2）用户信息表

此表用途是存储用户注册登录信息。表中有 id、Username、Password 三个字段名，id 为主键，属性自动递增，表示注册成功的用户编号，Username 表示注册时的用户名，Password 表示用户注册时的密码。如表 2 所示。

表 2 用户信息表

字段名	类型	长度	小数	不是 null	主键	内容
id	int	11	0	是	是	序号
Username	varchar	1024	0	否	否	用户名
Password	varchar	1024	0	否	否	密码

4 蒙古文语料库管理系统

为了有效地使用和管理所构建的蒙古文语料库，本文设计并实现了一个语料库管理系统。

4.1 系统设计

本系统主要由用户管理模块和语料库管理模块组成。其中用户管理模块主要包括两个子模块：登录和注册。语料库管理模块由查询统计、删除、添加、修改四部分组成。如图 2 所示为本系统功能结构图。

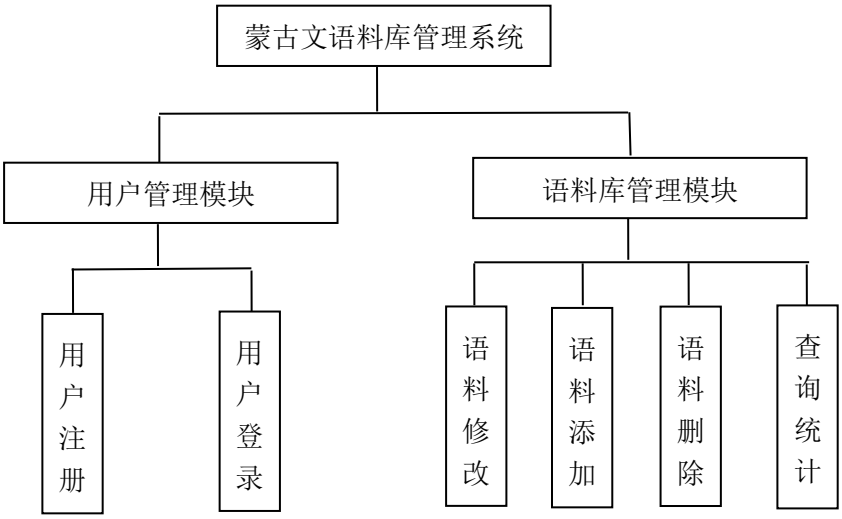

图 2 系统功能结构图

用户管理模块的“登录”功能，主要是根据用户输入的登录信息，到数据库中核对用户输入的账号密码是否正确。“注册”功能是将用户输入的登录信息存入到 user 数据表中。

蒙古文语料库管理平台对于通过注册、登录的用户，对语料库进行管理操作，包括添加，删除，更新，查询统计四种操作。其中“添加”功能用来对蒙古文文本数据库中添加新的文本内容。“修改”功能是对数据库中的蒙古文文本内容进行修改。“删除”功能则是用来删除数据库中选中的内容。“查询统计”功能是利用关键词查询来检索带有关键词的语句，并统计出带有关键词的语句总数。

4.2 系统功能实现

本系统利用的是 Python 自带的微型 Flask 框架建立了搜索引擎页面，Flask 是一个轻量级 Web 应用框架。然后，利用 Python 语言将数据库与搜索引擎进行了连接，并访问蒙古文语料库。

4.2.1 用户管理模块

用户管理模块分为用户注册和用户登录两个子模块，下面对这两个功能模块进行详细的介绍。

(1)用户注册

本系统注册模块需要用户在登录框内输入用户名、密码，随后前端会将登录表单发送给后台，后台会在 test 数据库中的 user 表中查找用户登录信息，最后

把结果返回给前端页面。如果注册成功显示“注册成功！请登录”，如未注册成功则显示“请填写此处”。用户注册成功时将用户名和密码两种登录信息存储在数据库中的用户数据模块 user 表中。详细步骤如下图 3 所示流程图。

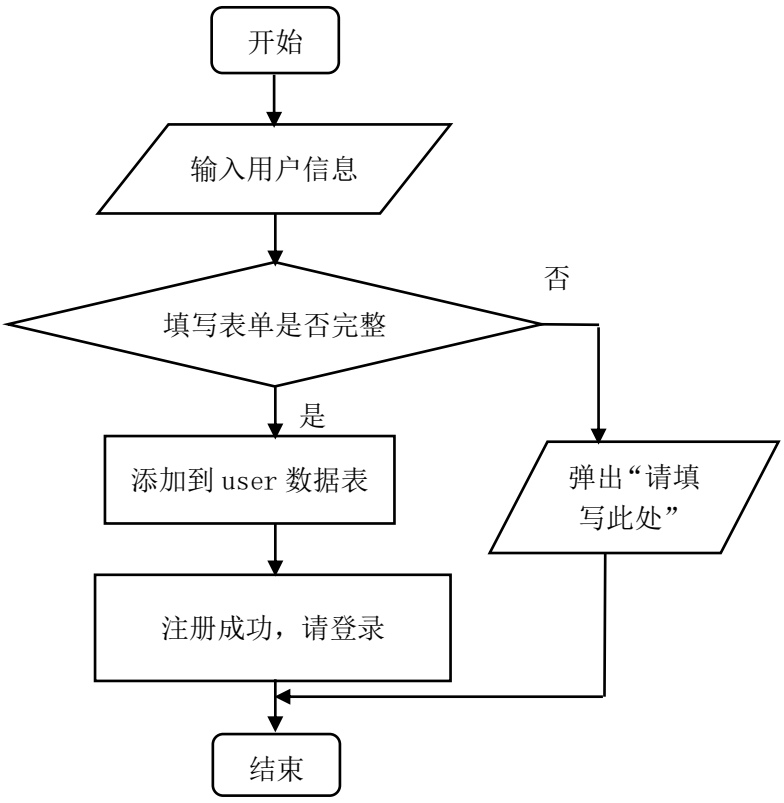

图 3 注册模块流程图

(2) 用户登录

用户需要填写注册时的用户名和密码进行登录，需要从 user 数据表中查询用户信息。当输入的用户名没有进行注册而登录时，系统将显示“未找到该用户！请先注册！”，点击注册显示“注册成功！请登录”时再进行登录操作。当用户输入的密码错误时显示“用户名或密码错误！请重试！”。当用户名和密码都正确时才可登录，页面跳转到语料库管理页面。详细步骤如下图 4 所示流程图.

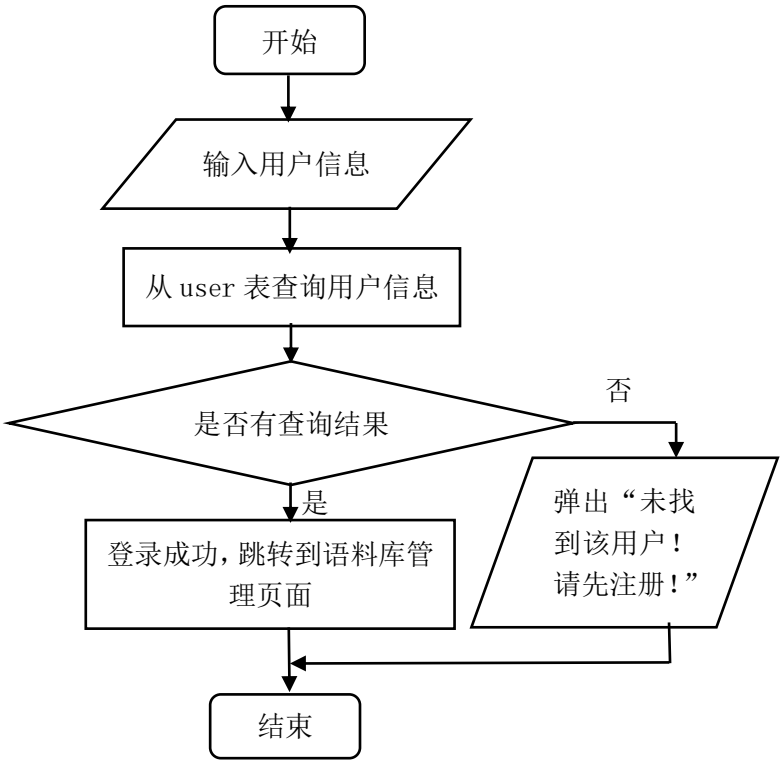

图 4 登录模块流程图

4.2.2 语料库管理模块

本模块对存入数据库的蒙古文文本数据进行“修改”，“添加”，“删除”，“查询统计”4 种操作。当用户登录成功，对文本进行相应的操作时，test 数据库中对应的 testtxt 数据表的文本数据也发生改变。

（1）语料添加：点击“添加”按钮，弹出“请填写”文本框，填写要添加的语料文本数据，在相应数据库中存入添加的文本内容。

（2）语料修改：点击“修改”按钮，可以对其文本内容直接进行修改，并按“确认”键可以看到修改后的内容。

（3）语料删除：点击“删除”按钮，把相应的文本内容从 testtxt 数据表中删除。

（4）查询统计：本模块是利用关键词进行查询并统计出带有关键词的语句总数。当用户输入被查询的关键词后，查询统计模块会遍历语料库的文本内容，显示出带有关键词的语句，并且用红色显示关键词。查找方式为顺序查找，字符串匹配模式为精确匹配，基本思想为根据给定的文本主串 T=T1...Tn 和用户输入的模式串 P=P1...Pm，在主串里查询与模式串相等的子串。如果匹配成功则表明

查询成功。即直接从数据库里面找到与关键词完全一致的目标功能。具体实现利用 SELECT id, case_process FROM `testtxt1` WHERE case_process LIKE "%{wd}%"查询语句对数据库内容进行关键词查询。详细步骤如图 5 所示。

统计算法的思想是当用户在搜索栏里输入查询文本内容，开始遍历查询结果的内容，当查询结果中有下一条语句时 count++（count 初始化为 0），并输出 count 的最后值，直到遍历完查询结果，显示共找到 count 条句子。

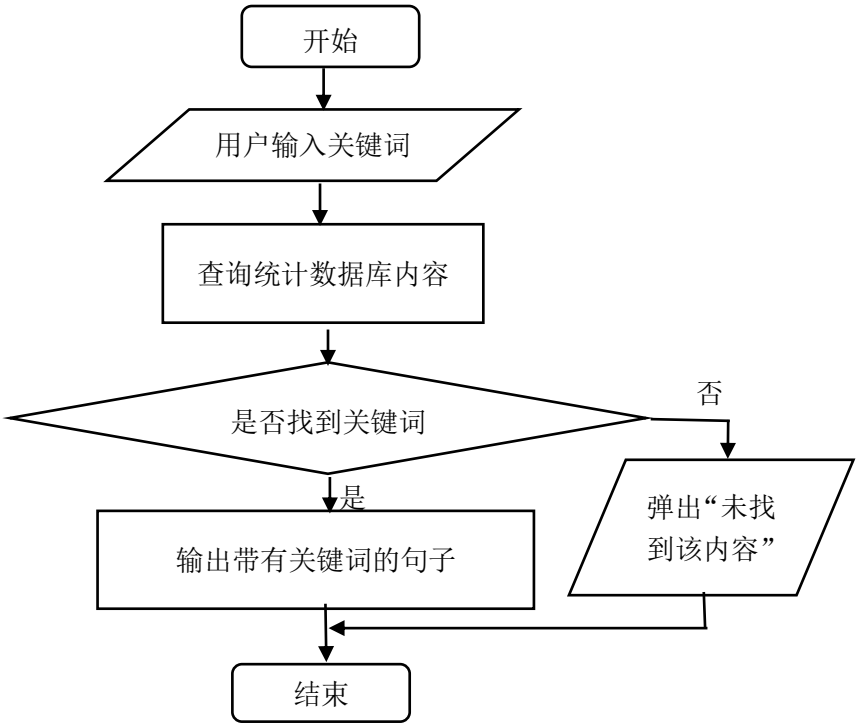

图 5 查询统计流程图

本文通过以上步骤，实现了本蒙古文语料库管理系统，系统运行的查询统计结果界面图如图 6 所示。

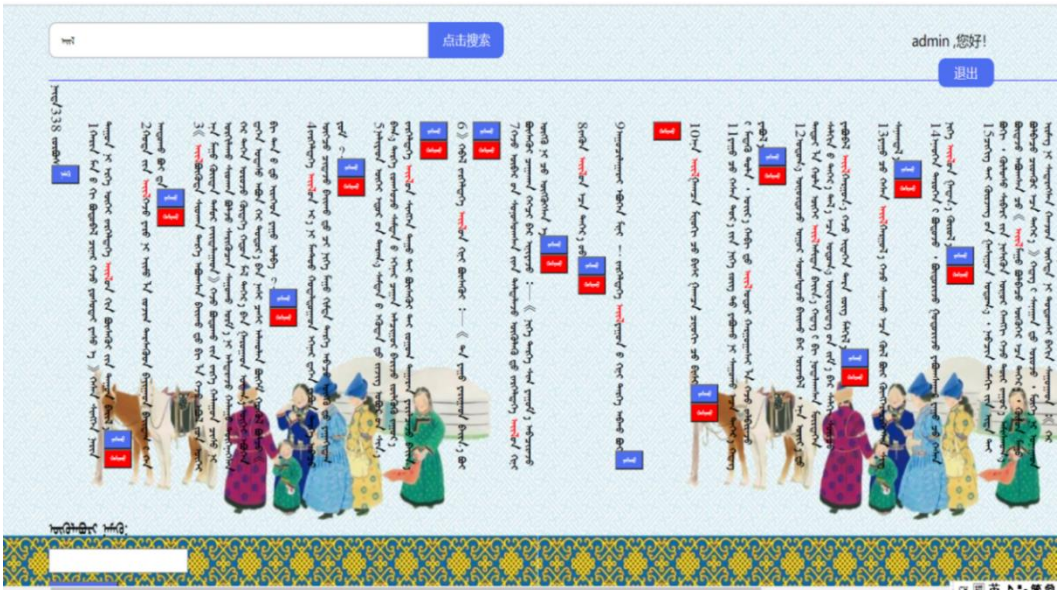

图 6 查询统计结果界面图

5 总结

本文主要根据特定网址爬取相关网站文本数据，对爬取后的文本进行预处理并存储到数据库中，设计和实现了蒙古文语料库管理系统，本系统主要有查询统计、插入、删除、添加等功能。

在本蒙古文语料库系统的创建过程中，初步了解了建立语料库的原理和步骤，这期间碰到的难题也不少，有些解决了，但有些还存在些许问题。比如语料库的内容不够丰富，不具分类等。因此，本蒙古文语料库系统需要改进的地方还有很多。

致谢：在本次毕业论文的写作过程中得到了王斯日古楞老师的精心指导，在此表示衷心的感谢。

参考文献

[1]斯日古楞. 《现代蒙古语语料库管理平台》建设[D]. 内蒙古:内蒙古大学,2010.

[2]哈斯.蒙古语语料库语言资源管理平台的设计与实现[J].内蒙古师范大学学报(自然科学汉文版),2008,37(06):743-745+749.

[3]龙梅. 基于蒙古语标注语料库的检索系统的设计与实现[D]. 内蒙古:内蒙古大学,2014.

[4]刘伟.语料库语言学发展现状与应用[J].山东电力高等专科学校学报,2009,12(02):42-44.

[5]潘永木梁 .语料库语言学的目的和方法 [J] .解放军外国语学院学报,2001.

Mongolian network data crawling and corpus management system

School of Computer Science and Technology 2017 NaHeYa 20172104256

Directed by WangSiriguleng Professor

Abstract This paper obtains Mongolian text data through web crawler technology, and preprocesses the crawled text content, including removing useless symbols, segmenting the text, and storing the segmented text into the database. Then, the Mongolian corpus management system is designed and implemented. The system mainly has the functions of adding, modifying, deleting, querying and statistics. This paper first introduces the methods of text data crawling and preprocessing, and then gives the design and implementation of corpus management system.

Keywords Mongolian data crawling；data processing；corpus management