

Development of Chinese Access
and Search System
for the Digital Silk Road Project

Tao Zhang
Department of Automation
Tsinghua University

Contents

- Introduction
- Chinese Access and Search System
- Search Algorithms
- Conclusions

Introduction

- Digital Silk Road Project
- Collaboration between Tsinghua University and National Institute of Informatics of Japan
- Assist Chinese people who have interests in the history of Silk Road to obtain the information about Silk Road
- Develop a Chinese access and search system
- Propose a search algorithm

Chinese Access and Search System

- Three approaches to locate the expected information from the Digital Silk Road System developed by NII.
 - Directly access the Chinese version of the Digital Silk Road System translated by Chinese people, which will be finished and installed in NII.
 - A Chinese system developed by us and installed in Tsinghua University, by which user can easily link to the expected information in the Digital Silk Road System.
 - A search tool developed for the Chinese system so that Chinese people can easily locate the information by using Chinese words.



[English](#)

デジタル・シルクロード

デジタル・シルクロード・プロジェクト (Digital Silk Road Project) は、情報学と人文学の融合に基づく文化遺産のデジタルアーカイブを構築する研究プロジェクトです。 [[もっと詳しく >>](#)]

新着情報

遷画へシルクロード

NEW サイトがオープンしました。ただしヘルプ等のページは後日に公開予定です。

2007年08月23日

- [「東洋文庫所蔵」図像史料マ](#)
[ルチメディアデータベース](#)
- [イラン・バムの城塞：地震を越えて残す記録](#)
- [遷画へシルクロード](#)
- [地図で探るシルクロード](#)
- [DSRイマジナリミュージアム](#)
- [シルクロード談叢](#)
- [写真でつなぐシルクロード](#)
- [シルクロード用語集](#)
- [デジタル・シルクロード・キッ](#)

Chinese Access and Search System

- A C++ program is developed for automatically translating the contents in English into Chinese.
 - Folder Part
 - File Part

Program and
Results

Chinese Access and Search System



Digital Archive of Toyo
Bunko Rare Books

数码版的东洋文库希见书

数码档案文件包含 53本珍贵书籍 (19 authors : 13,973 pages) with narratives about Silk Roads. [\[Read more...\]](#)

当前位置 > 首页

检索网站内容:

Search

重置

提交

Introduction to Website

Link for
contents

网站介绍:

.1 印度和高地亚洲--地理部分 XII-4-2

.2 走向中国 III-2-F-b-2



Chinese Access and Search System

数码版东洋文库希见书

[数码丝绸之路项目](#) > [东洋文库档案文件](#) > [III-2-F-b-2](#)

ja en



标题 走向中国
副题 中国老式布告收藏
作者 [\[Yule, Henry 爵士\]](#)
梗概 《走向中国》是一个早期的关于中国的评论合集，由Yule整理编辑而成。曾获皇家地
出版日期 1866
出版地点 英国 伦敦
卷数 2卷
语言 [英语](#)

[III-2-F-b-2](#)

(a)

Digital Archive of Toyo Bunko Rare Books

[Digital Silk Roads Project](#) > [Toyo Bunko Archive](#) > [III-2-F-b-2](#)

ja en



Title Cathay and the way thither
Subtitle being a collection of medieval notices of China
Creator [\[Yule, Sir Henry\]](#)
Description "Cathay and the way thither" is a collection of comments on China from early days
compiled and edited by Yule. It was awarded a gold medal from the Royal Geogra
Society.
Year of Publication 1866
Location of Publication England / London
Volume Information 2 Volumes
Language [English](#)

[III-2-F-b-2](#)

(b)

Search Algorithms

- Basically, search algorithm includes three parts:
 - URL catching
 - Web page pre-processing
 - Query system

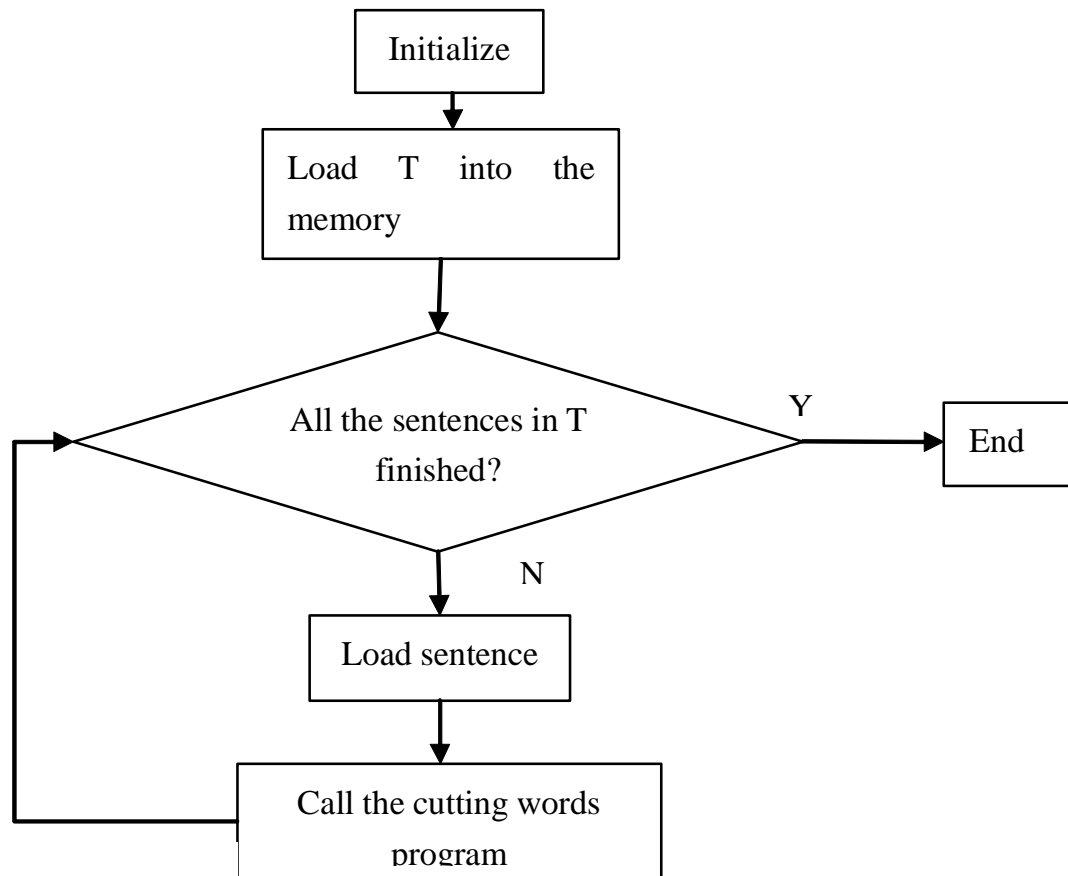
URL Catching

- Procedure:
 - Analyze the server address from URL pool
 - Set up the connection
 - Send the request and accept the data
 - Save the data into database
- Solve two problems:
 - Searching one same web page several times
 - Storage format

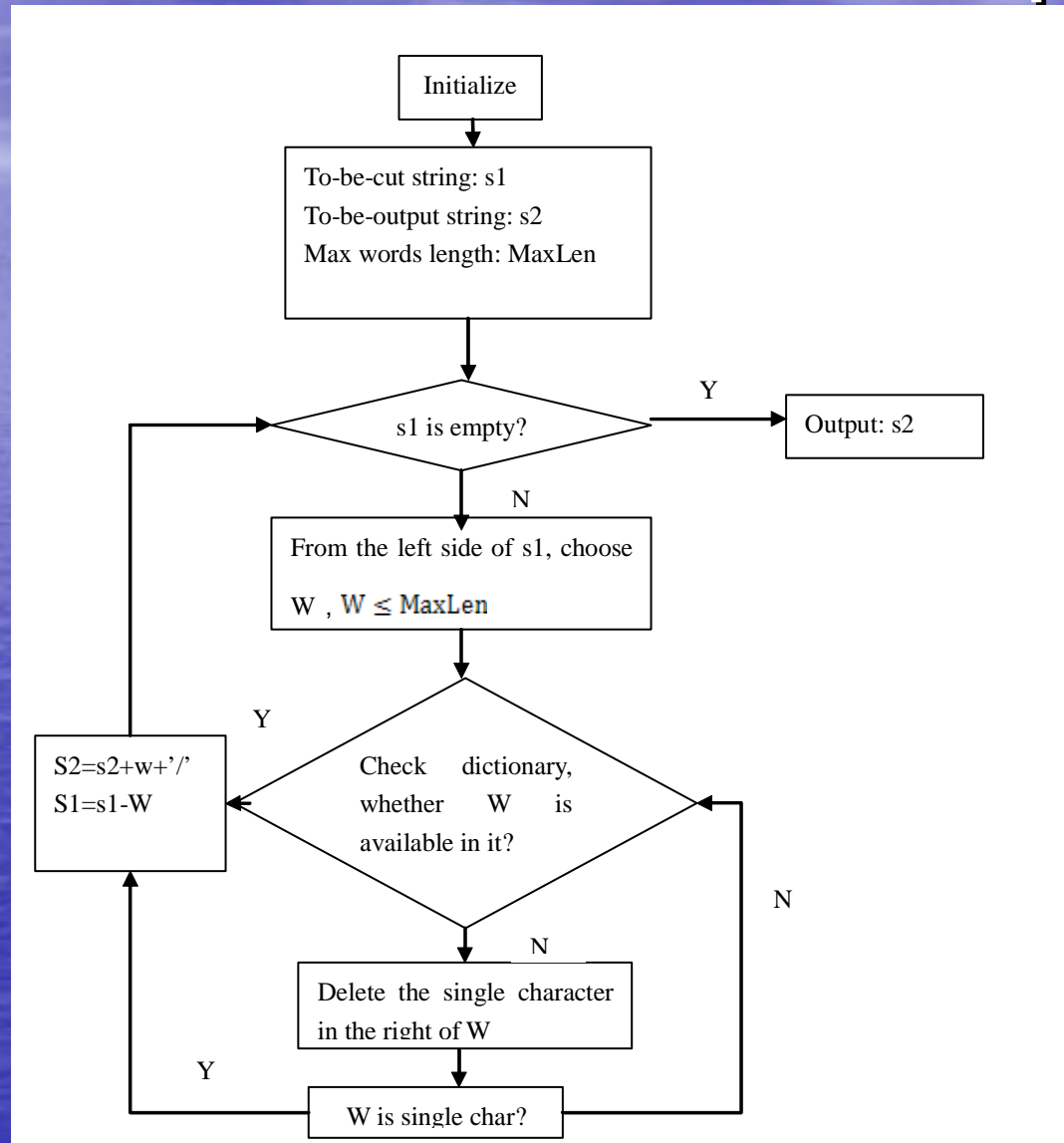
Web page pre-processing

- Set up index database
- Chinese words cutting
 - Mechanical Cutting Words Method (based on string matching)
 - Maximum Match Based Approach
- Set up reverse index table
 - General index table
 - Reverse index table

Maximum Match Based Approach



Maximum Match Based Approach



Query system

- Server gets user request
- Cut the user request sentences into the combination of several words
- For each word the relative web pages will be got from the reverse index file and an intersection among the result will be made
- Feed back the results to user

Conclusions

- Develop a Chinese access and search system
- Propose a searching method for Chinese Access and Search System
- Improve it for new information
- Provide real service for Chinese



Thank You !