

基于主题抽取的网络信息检索系统

——技术和实现

- 罗强
- 华南理工大学
- qluo163@163.com

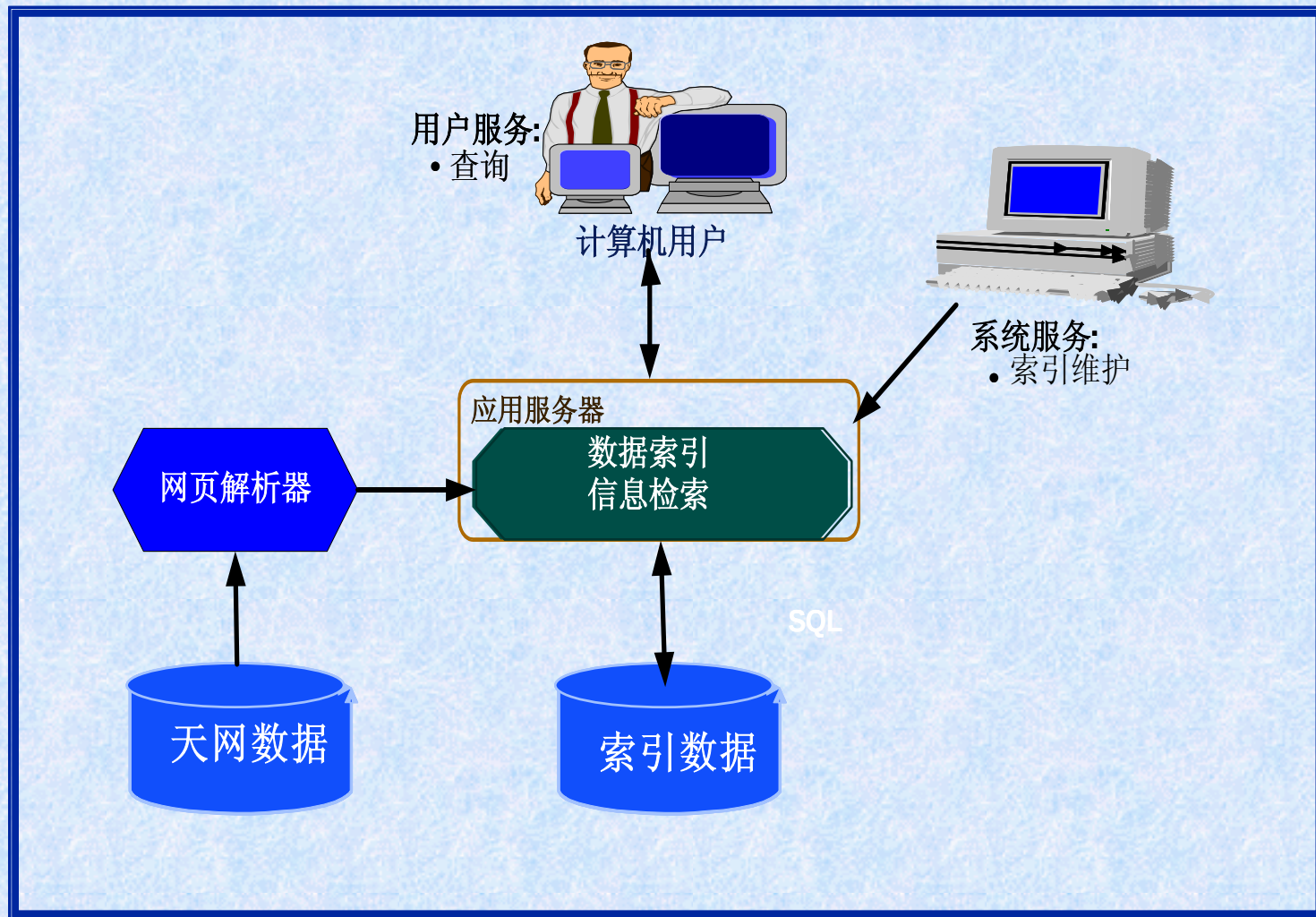
检索技术存在的问题

- 网络资源庞大而易变
- 信息系统难以获取主题信息需求
- 难以将信息需求转化为查询表示

报告内容

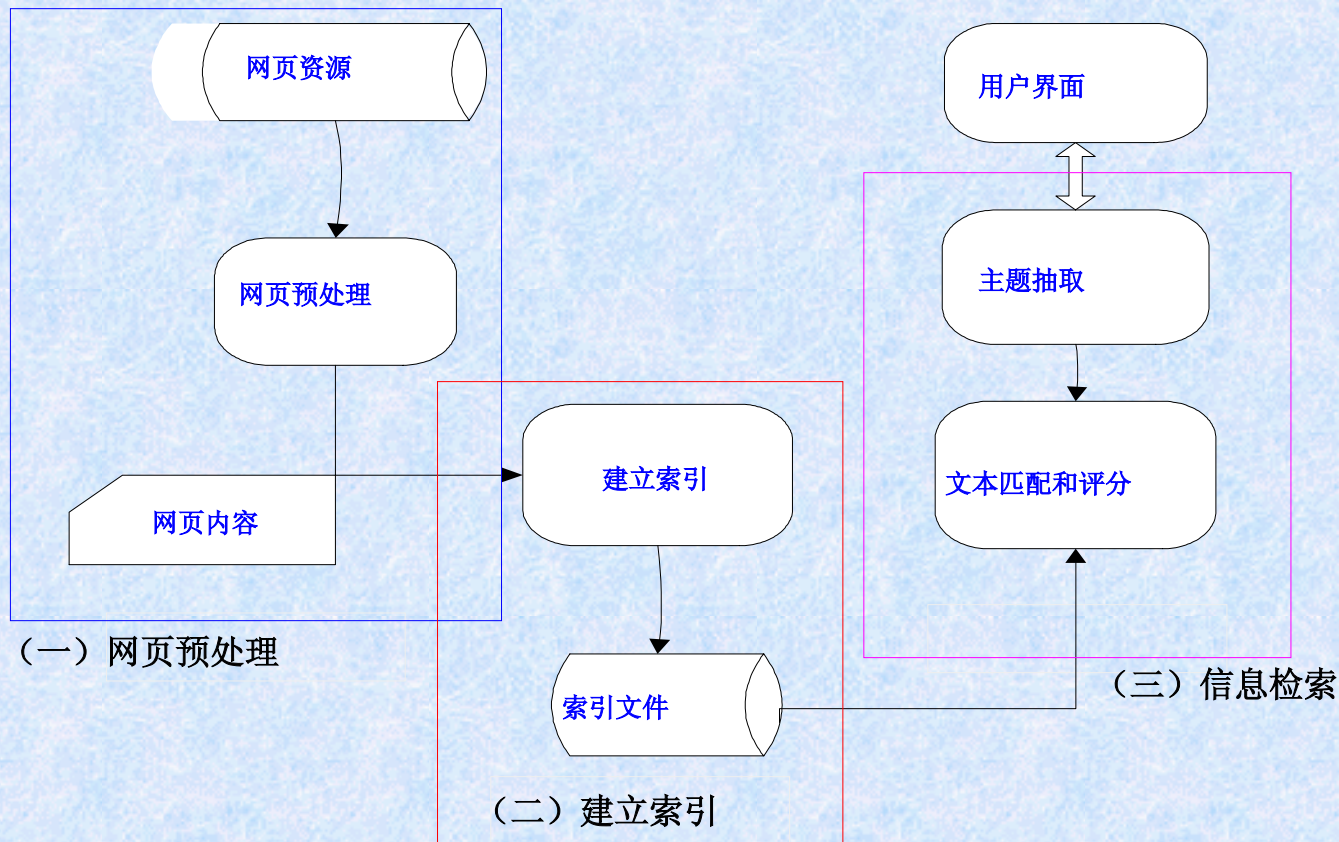
- 1、系统架构
- 2、系统组成
 - 数据预处理
 - 建立索引
 - 信息检索
- 3、关键技术介绍
- 4、总结

1、信息检索系统架构



2、系统组成

——信息搜索模块



2.1、网页预处理

- 网页内容划分

- 划分为两个域，标题（Title）域和内容（Content）域

- 网页解析

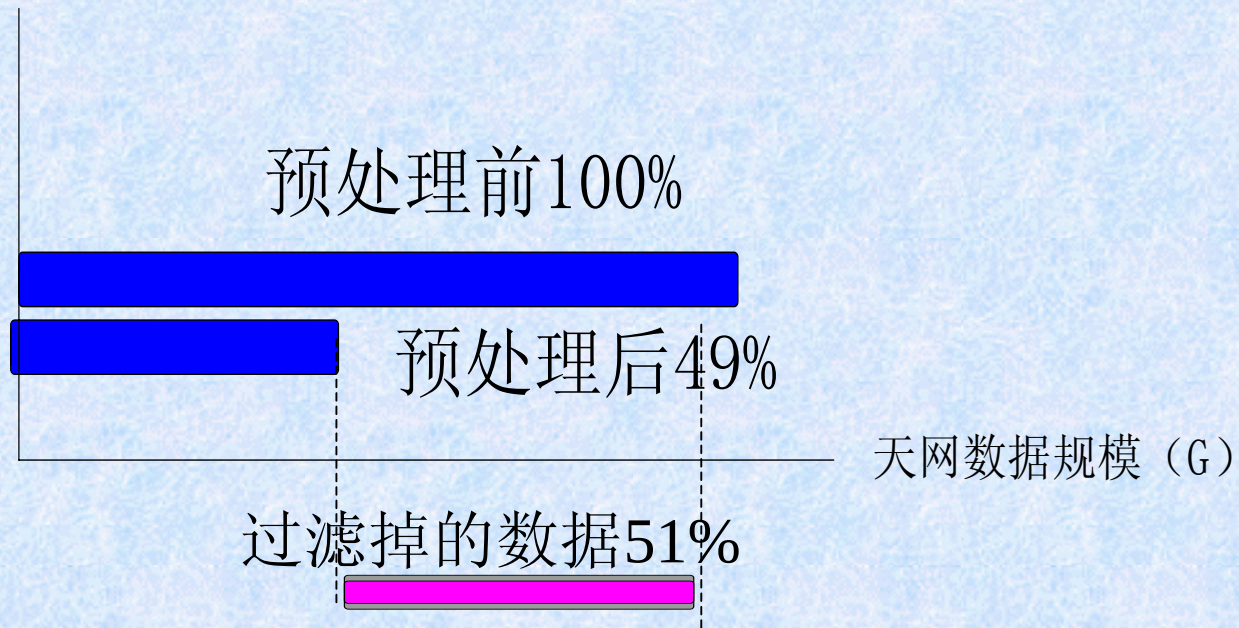
- 遍历网页的各标记结点
 - 重点考虑的标记有
 - `<url><title><table>`
 - 不考虑的标记：与图形、脚本、链接及链接文本有关的标记
 - 如 `<script><a>`等

- 文本合并

- 合并遍历网页所获取的文本，分类存放到标题（Title）和内容（Content）域中

- 生成存储网页内容的XML文件

天网数据预处理情况



jsp/asp脚本，图片，链接和链接文本，网页
标记，特殊格式的文件
(.exe,.com,.jar,.zip,.pdf,.doc等后缀结尾的文件)

2.2、建立索引

- 索引结构选择
 - 采用倒排索引树
 - 比较成熟、创建和查找速度快，支持索引的增量添加、易于实现
- 索引的域
 - 构成
 - 标题域（**Title**）
 - 内容域（**content**）

天网数据创建索引的情况

内容是否保存	索引规模 (M) /G	索引/天网数 据	处理速度 (M/S)
是 (未采用)	≈ 600	58.6%	0.2-0.3
否 (采用)	15-20	1.8%	0.4-0.5

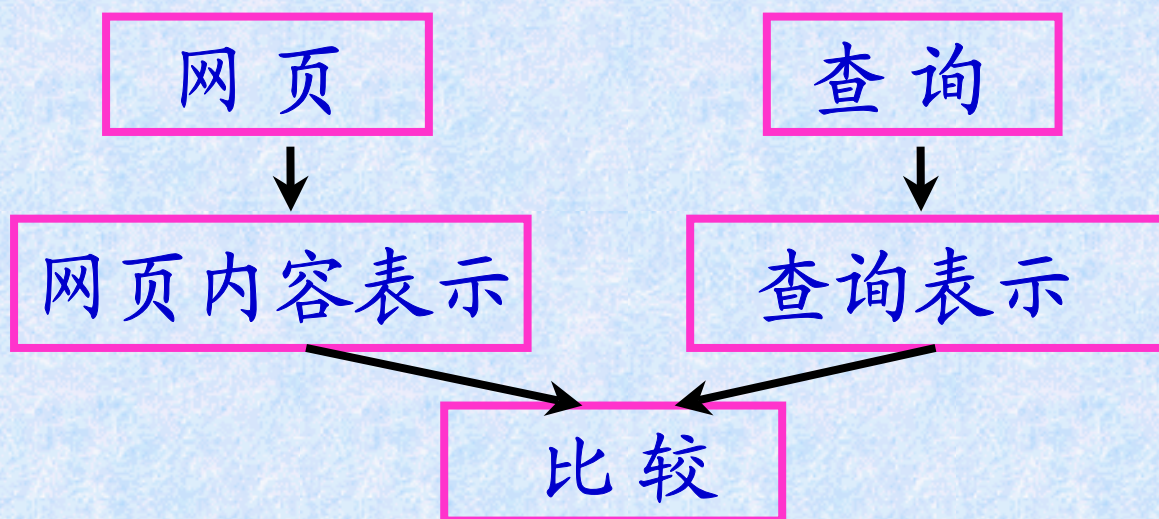
处理数据总量：**70G**

索引大小：**约 1G**

注：**内容是否保存**是指网页内容是否保存到索引文件中

2.3、信息检索

信息检索的基本原理



- 基于网页的内容描述，产生网页的文本表示
- 针对主题的信息需求，产生信息需求表示
- 比较上述两种表示，选择与主题信息需求最接近的文本

信息检索的组成部分

- 文本预处理
- 查询操作
- 结果排序

2.3.1 文本预处理

- 文本词集
 - 通用词集（约8万余条）
 - 专用词集（约1千条）
- 文本词法分析
 - 分词
 - 专用词、名字实体标记
 - 数字，字母，特殊符号标记

2.3.2 查询操作

- 获取主题的信息需求
 - 主题类型
 - 主题的中心成份
- 转换主题信息需求为查询表示
- 搜索索引文件，对候选文本评分
 - 统计主题成份的词频，获取每个词的权重（**TFIDF**）
 - 计算候选文本与主题的相关度(**VSM**)

2.3.3结果排序

- 对每一条主题，进行三次查询
 - 标题（**title**）域
 - 内容（**content**）域
 - 标题（**title**）&内容(**content**)域
- 对上述查询进行综合评分
- 返回评分最高的前 ‘N’ 个文本

3、主要技术介绍

基于主题抽取的信息查询技术

3.1 此项技术应考虑的问题

- 主题信息需求的范围有哪些，它们是如何组织的？
- 如何表示和获取主题的信息需求？
 - 语法角度：信息需求 = 主题中心成份 + 约束关系
- 如何从主题的信息需求生成查询表示？

3.2主题抽取系统的定义和任务

- 定义

主题抽取系统是一个主题信息抽取模块，它运用事先构造的规则集在每一个处理步中抛弃一些无关的信息，从而获取主题的关键内容。

- 任务

- 分析主题信息描述
- 主题分类
- 主题成份关系获取

3.3主题信息抽取模块构成

- 规则库

- 获取主题规则： 手工添加
- 规则形式： 主题模式（模板）

- 信息分析模块

- 输入： 主题或词条序列
- 信息添加： 主题成份间的依存关系
- 信息丢失： 无关信息（如停用词）
- 输出： 主题的关键内容

3.4 实现

- 预处理
 - 去掉停用词，转换为更小的词条序列
 - 分词、词性标注、名称实体判别 (NE)
 - 浅层句法分析（文本组块）
- 主题分析
 - 主题分类
 - 约束关系获取
- 主题模式匹配
 - 主题中心成份获取

4、小 结

- 主题抽取技术是基于特定的领域和假设
- 通过分析文本获取感兴趣的特定信息
- 转换主题信息需求为查询表示

谢 谢 大 家！

Q&A