

The background of the slide is a deep blue with a fine, woven texture. There are two rectangular areas with a lighter blue, abstract, and somewhat pixelated pattern: one in the top-left corner and another in the bottom-right corner.

内容计算与信息服务

---富士通研究开发中心介绍

2004年8月27日

- ❑ 研究目标
- ❑ 研究内容
- ❑ 成果和产品

- 研发以多媒体处理为核心的信息处理技术, 特别是基于内容的知识收集、分析、再利用技术
 - 汉语处理工具
 - 图像内容检索、识别
 - 语音识别、合成

互联网时代信息的特点



- ❑ 信息量大
- ❑ 变化快
- ❑ 多信息融合
 - 结构化、半结构化、非结构化信息的融合
 - 照片→APS照片→数码照片→内容数码照片
 - 文本、音频、视频信息以及传统信息与结构化信息之间的融合

基于内容的信息处理

根据客户的需要处理信息

- 没有必要知道的信息过多，注重提供准确，明确，适当的信息
- 提供帮助性信息，辅助用户信息获得

有效率的收集
→ 知识性Web Crawler

整理，体系化→
自动分类，信息抽取

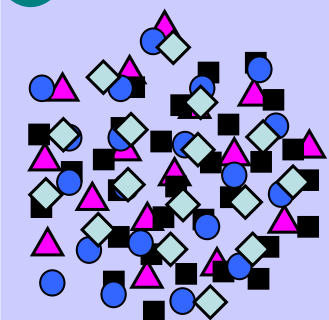
信息活用→
信息分析，知识检索，
信息监控

Internet

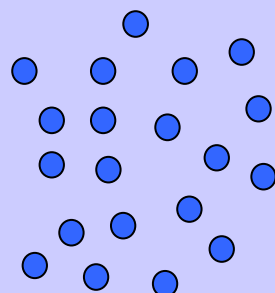
1. 收集

2. 整理

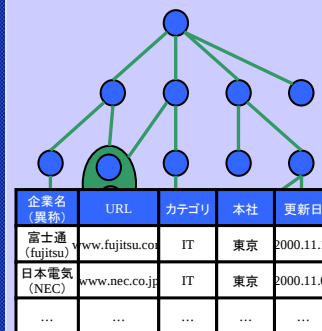
3. 活用



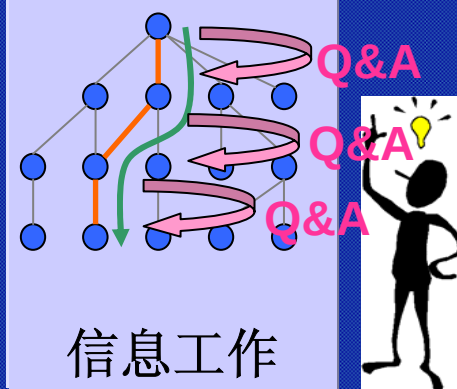
大规模多样化



特定优良信息



结构化信息



信息工作

□ 汉语语言处理工具的整備

- 专有名词识别，专门用语分析
- 基于因特网的语言知识获得

□ 多媒体信息处理技术的研发

- 图像检索（Web、设计图、3 D图像）
- O C R（汉语手写地址识别、视频图像中的文字识别）
- 语音处理（汉语语音合成、汉语语音识别应用）

□ 应用系统

- 基于内容的知识检索、管理
 - 多媒体检索、识别，中国企业信息分析
- 日中辅助翻译, 汉语语言辅助学习工具

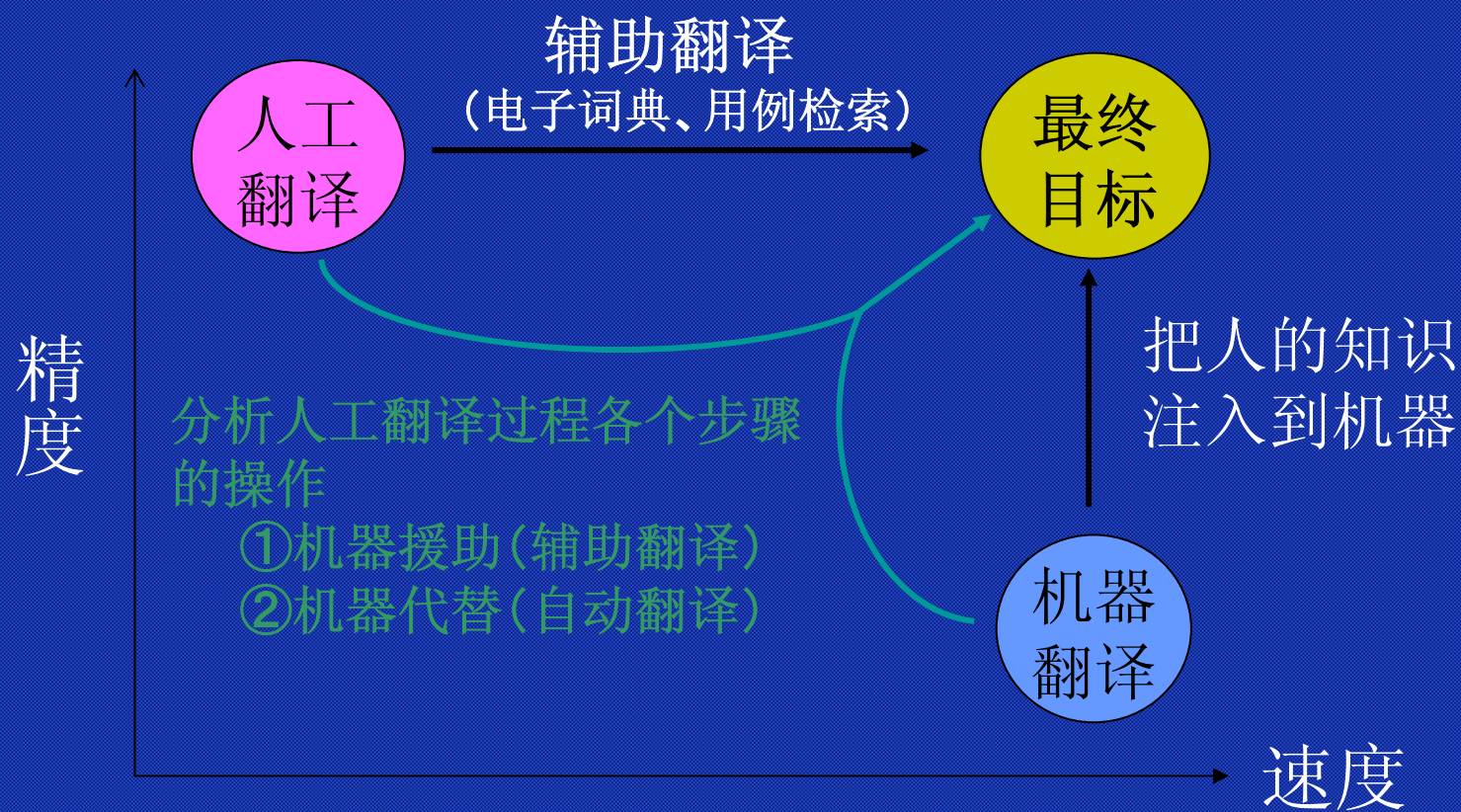
成果和产品介绍

- ❑ 与北京大学、人民日报合作开发
- ❑ 1998年全年，约2000万字
- ❑ 已完成并公开的国内最大的分词和词性标注资源

- ❑ 汉语分析（分词，词性标注，基本短语分析）
- ❑ 汉语专有名词（人名，地名，组织名）识别
- ❑ 汉语读音标注（多音字，数字读音的考虑）
- ❑ 面向因特网的汉语新词发现
- ❑ 面向因特网的中英，日英互译语词典的抽取
- ❑ 专门用语分析（英语文献的蛋白质名的识别）

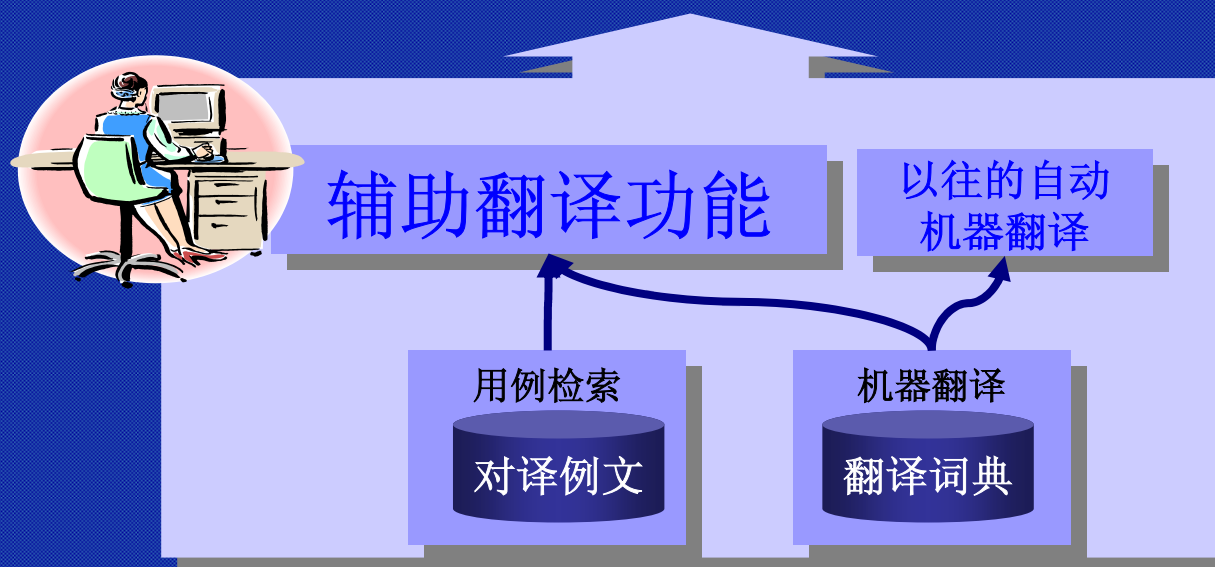
辅助机器翻译系统

高精度、高速的翻译目标



机器翻译和实例检索结合

- ❑ 完全的自动翻译
 - 高速自动生成译文
 - 可以达到易懂，但不充分，语言不流畅
- ❑ 辅助机器翻译
 - 由人最终确认生成译文，从而提高效率、质量
 - 机器翻译和实例检索结合，保证翻译一致性，高效的再利用企业内部的翻译词典与例文,实现翻译知识的积累与共享
 - 可以完成译文质量要求高的翻译任务



机器翻译与实例检索界面（英日）

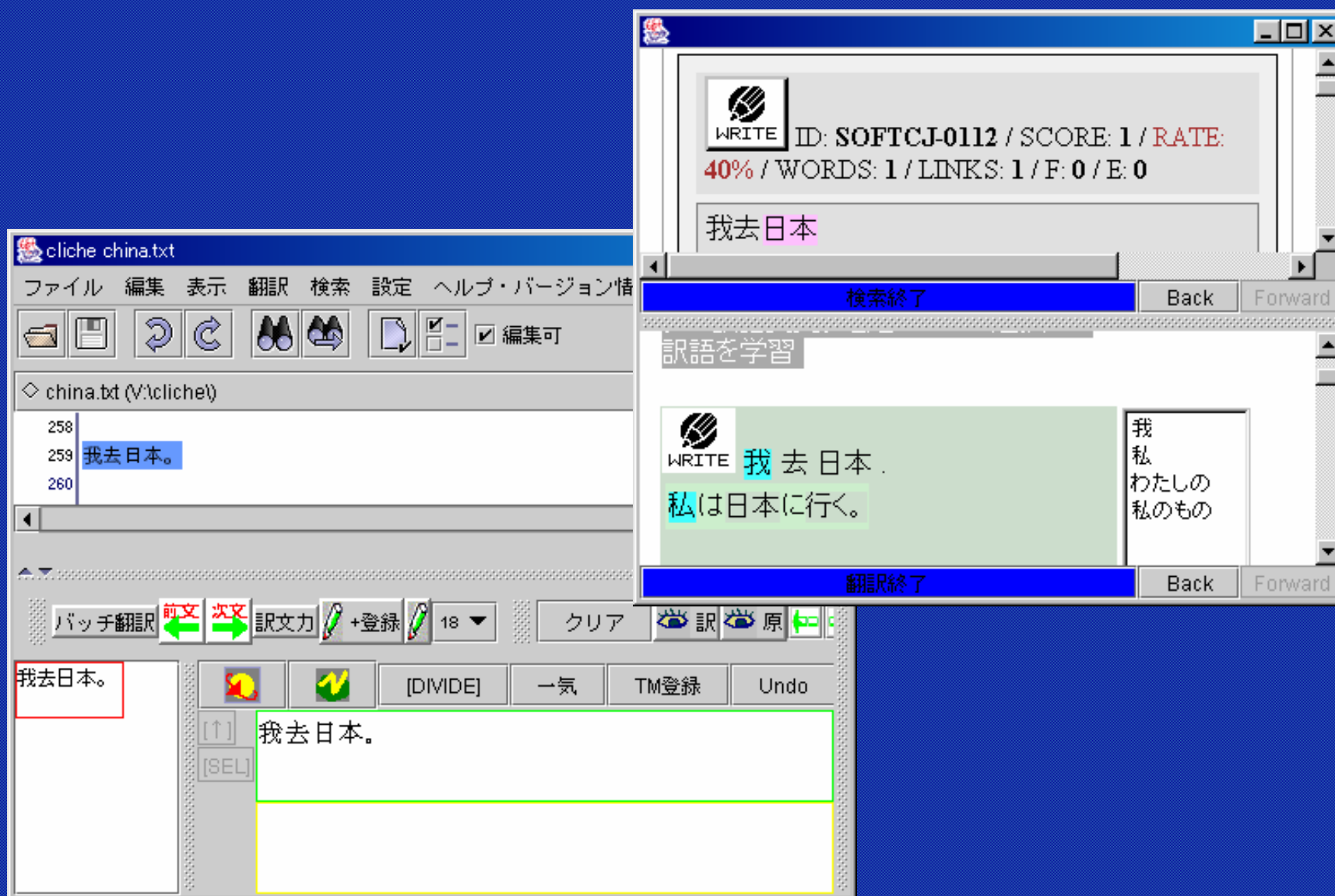
The screenshot displays the ClicheEditor interface for machine translation. The top menu bar includes options like 'ファイル' (File), '編集' (Edit), '表示' (View), '翻訳' (Translate), '検索' (Search), '設定' (Settings), and 'ヘルプ・バージョン情報' (Help/Version Information). The main workspace is divided into three sections:

- HTML Source Code:** The top-left pane shows the HTML source code for 'Agreement1.html'. It includes tags for paragraphs, strong text, and alignment. A line of code is highlighted: `<DD>Noting that nothing in this Agreement or its implementation shall be construed as prejudicing either Party's position on issues of competition law jurisdiction in the international context, </DD></DL>`.
- Translated Text Preview:** The middle-left pane shows a preview of the translated text: "Noting that nothing in this Agreement or its implementation shall be construed as prejudicing either Party's position on issues of competition law jurisdiction in the international context,".
- Translation Explanation:** The bottom-left pane provides a detailed explanation of the translation. It states: "本契約のいかなる部分も、決して契約違反より生ずるいずれかの当事者の相手方に対する権利及び償還請求権の制限として解釈されないものとする。" (No part of this agreement shall be interpreted as a restriction on the rights and claims for reimbursement of any party against the other party, arising from a breach of the agreement).

The right-hand pane displays the search results for the selected text. It shows the 'WRITE' button, the ID 'KEI-09A-0608', the score 'SCORE: 24 / RATE: 43%', and the word count 'WORDS: 11 / LINKS: 2'. Below this, the translated text is shown in a box, and a detailed explanation of the translation process is provided, including a note about the use of the machine translation results and the importance of understanding the context.

The bottom status bar shows the current page number '44,68' and the percentage '12%'. The taskbar at the very bottom includes icons for various applications and the system clock.

机器翻译与实例检索界面（汉日）



网上歌词自动检测系统

提高监视业务的效率



违法性
判别

TerraCrawler

- 网站经营者
- 邮件地址
- 曲名
- 歌手名等

信息抽取

自动分类
(音乐网页判别)

信息收集

抽取候选
(生成列表)

管理者、经营者
相应处理

检索引擎

因特网

合法・违法网站

向网站经营者
发出警告邮件

□ 应用

企业调查、业界表现，股价波动的根据

□ 概要

通过因特网自动收集中国企业相关信息的，对重要信息抽取整理，提供一般企业数据库里没有的动态更新信息。

□ 成果

- 从1.3亿URL(7千万Web网页)抽取中国企业的信息
25.2万条
- 当日信息、过去信息的统计（热点企业，热点产品等），分类信息等的提示

“富士通中国”的信息



企业动态

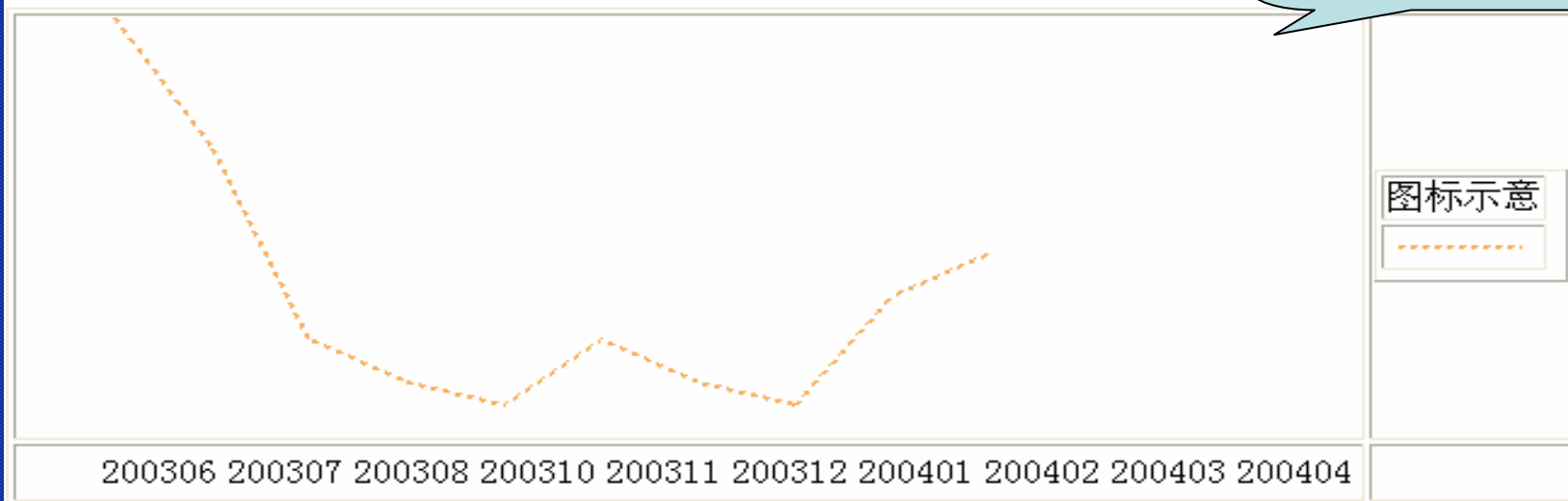
[富士通（中国）有限公司](#) (★★★★★) 知名度评价

企业新闻 (共有621条) [更多新闻...](#)

2004-04-20 TOM科技频道 [《商业周刊》：富士通日本一芯片马拉松新选手_TOM科技](#)
2004-04-14 TOM科技频道 [富士通、三星和爱普生纷纷建立显示面板工厂_TOM科技](#)
2004-04-12 TOM科技频道 [南京富士通打印机将以系列活动喜迎建厂十周年_TOM科技](#)

活动态势

来源 分类 时间单位



最近的新闻报道

活动活跃度的推移

“富士通中国”的信息



200306 200307 200308 200310 200311 200312 200401 200402 200403 200404

相关网站

其它频道:

关联网站: [Fujitsu Hong Kong Limited](#) [江苏富士通通信技术有限公司](#) [Hong Kong](#) [更多...](#)
友好网站: [通力互动](#) [赞华电子系统有限公司](#) [town@nifty](#) [更多...](#)
关注企业: [日立有限公司](#) [NEC公司](#) [美国Informix公司](#) [更多...](#)

相关公司，友好公司（合作等），竞争公司

分类

行业:

地区: [北京](#) [北京](#)

地域分类

业务类别:

投资类别:

基本信息

主页:	http://www.fujitsu.com.cn/	Email:	
电话		传真	
地址		邮编	

富士通产品关注度的变化

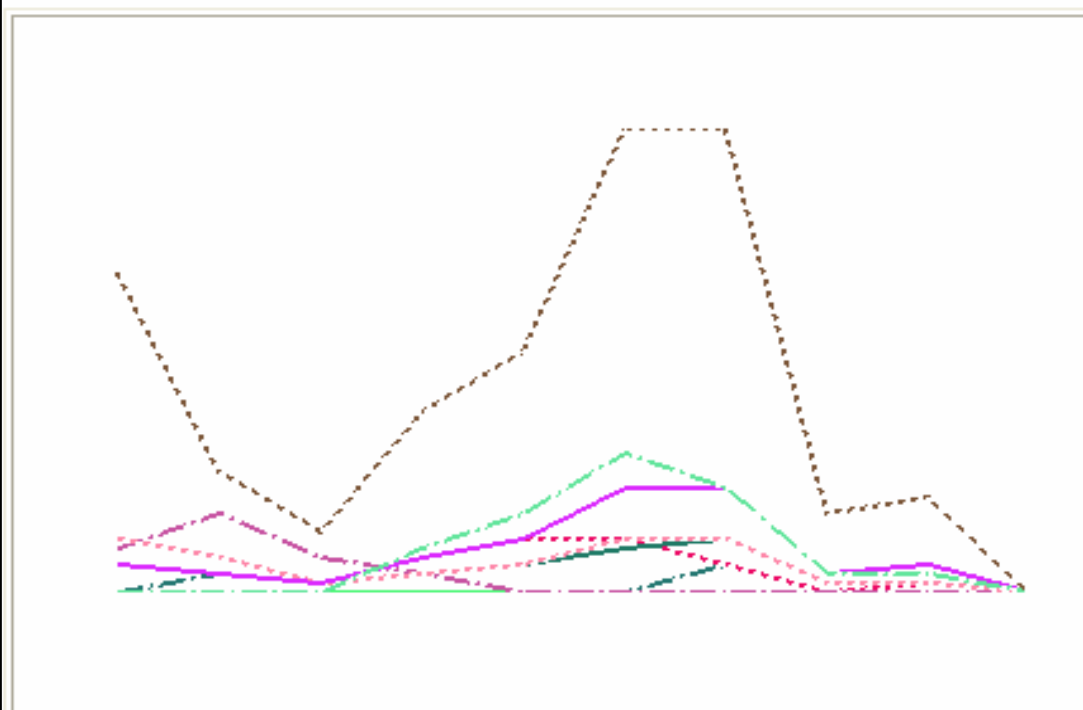


活动态势

来源

分类

时间单位



图标示意

-----	笔记本 (328)
-----	硬盘 (71)
-----	掌上电脑 (61)
-----	CPU (57)
-----	无线网 (44)
-----	液晶显示器 (40)
-----	手机 (27)
-----	服务器 (9)
-----	打印机 (1)
-----	音箱 (1)

2003年26周 2003年27周 2003年28周 2003年29周 2003年30周 2003年31周 2003年32周 2003年33周 2003年34周 2003年46周

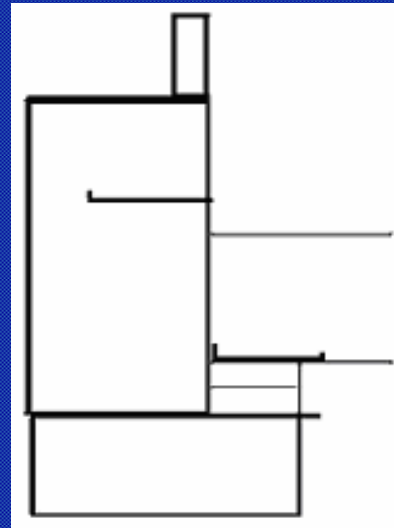
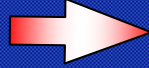
蛋白质名识别系统

- ❑ 背景：日本遗传学研究所与富士通共同开发世界最
高速的新一代生物数据库。数据库需存贮蛋白质名
称及关系
- ❑ 利用自然语言处理技术，从生物学文献中抽取蛋白
质名称及其关系

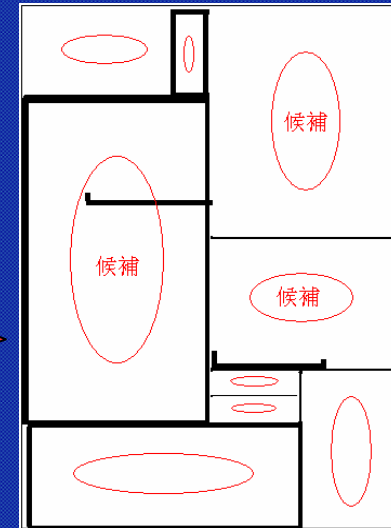
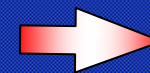
人民日报新闻图片检索系统



版面
分析



报导
分割



- ❑ 以新闻中的作者、日期、地点、事件等关键词索引
- ❑ 图文并茂地检索人民日报中的新闻
 - 通过版面分析技术抽取一篇篇单独的报导
 - 利用字体的大小、颜色及在报导中的位置识别标题
 - 根据图片与说明的相对位置、分割符、关键词来提取图片的说明

人民日报新闻图片索引示例

ReportTitle: 上海磁浮示范运营线举行通车典礼

Author: 钱彤

Date: 12-31

Content: 朱镕基和德国总理施罗德出席新华社上海12月31日电(记者钱彤、倪四义、吴宇)世界第一条商业化运营的高速磁浮交通系统——上海磁浮示范运营线31日举行典礼。国务院总理朱镕基和正在中国访问的德国总理施罗德出席典礼并发表讲话。上午10时,朱镕基总理和夫人劳安、施罗德总理来到装扮一新的示范运营线始发站——龙阳路站。两国总理为试运行通车剪彩后,同乘列车前往终点站——上海浦东国际机场站。10时10分,绿黄白相间的流线型磁浮列车如御风而行,行间达到设计最高运行速度每小时430公里,全程30公里只用了不到8分钟时间。中德两国总理和双方随行人员热烈鼓掌,对通车成功表示祝贺。随后,朱镕基和施罗德参观了浦东国际机场站站台,向车站工作人员频频挥手致意。再次登车返回龙阳路站后,两国总理一同参观了示范运营线总控制室,询问了磁浮交通系统建设和试运行的有关情况。11时10分,上海磁浮示范运营线典礼在上海国际新闻中心举行。朱镕基和施罗德发表讲话。(下转第二版)

Keyword: 龙阳路 速度 机场 吴宇 德国 上海 国务院 新华社 车站 典礼 终点 倪四义 中德 劳安 钱彤 夫人 施罗德 浦东 列车 磁浮 随后



Author: 马占成

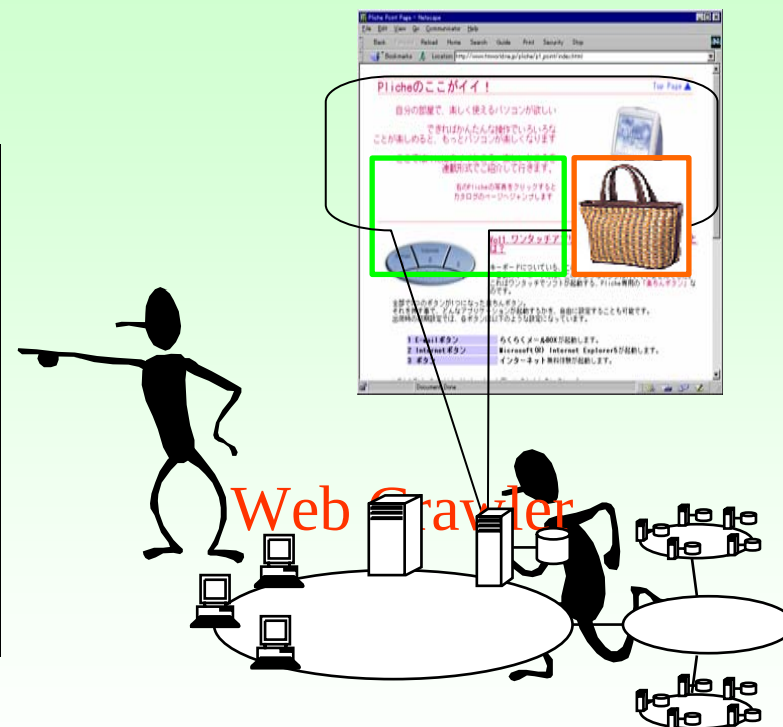
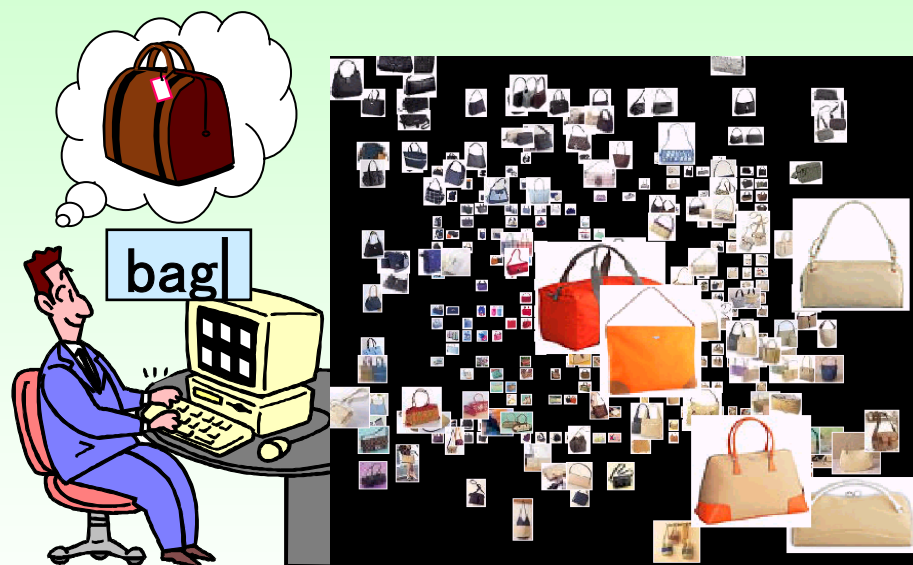
Date: 12-31

Keyword: 机场 德国 上海 国务院 新华社 马占成 终点 施罗德 浦东 列车

Explanation: 12月31日,上海磁浮示范运营线试运行通车。国务院总理朱镕基和德国总理施罗德为试运行通车剪彩后,一同乘坐磁浮列车前往示范运营线终点站——浦东国际机场站。新华社记者马占成摄

跨媒体网络信息检索系统

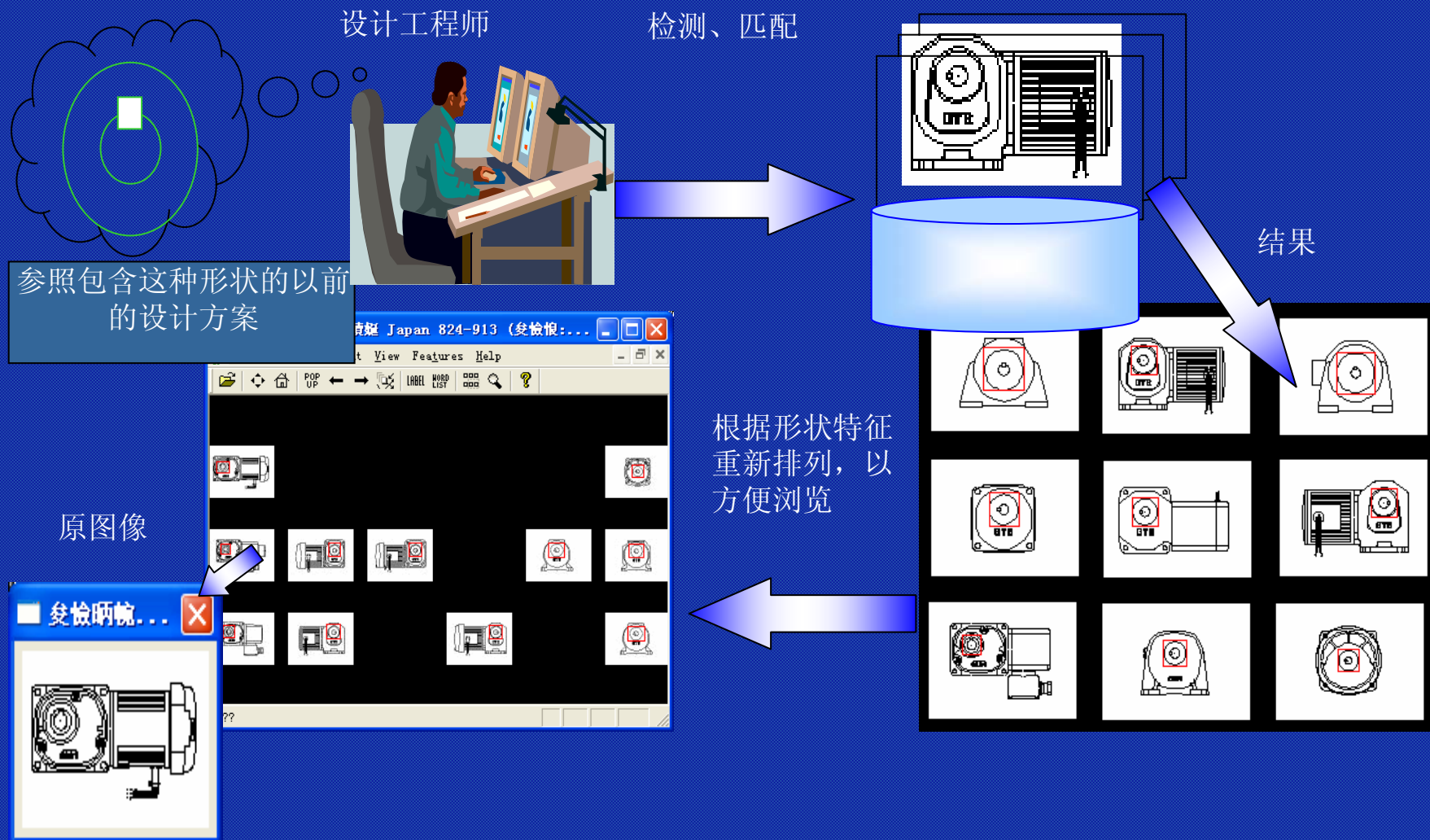
- 利用文本的图像检索
- 利用图像类似度的三维显示



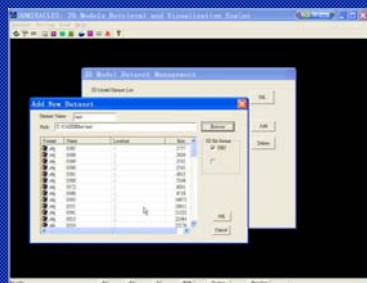
应用：设计图检索、零件检索

设计图检索系统

检索需要的工程设计图



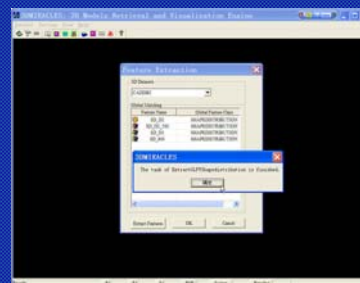
3D模型内容检索系统



模型数据库管理



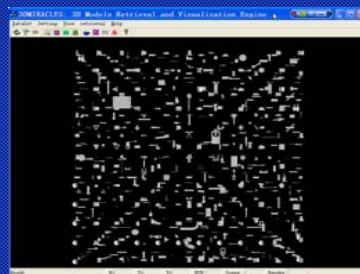
三维形状特征管理



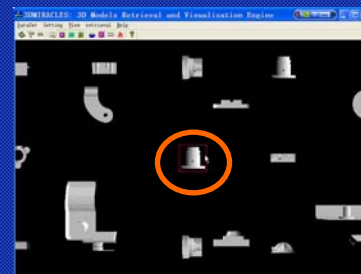
特征数据提取

系统数据管理

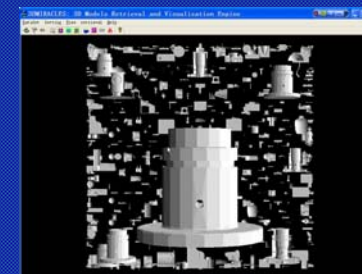
检索查询交互



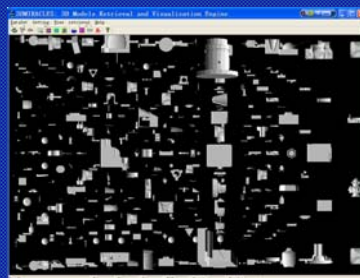
数据库浏览



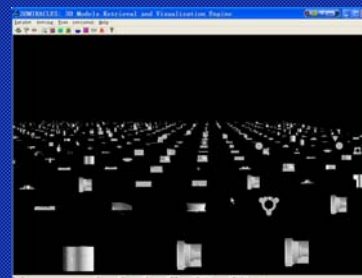
查询对象选择



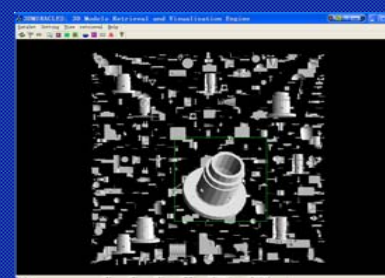
检索结果显示



模型库场景漫游



三维排列模式



三维操纵观察

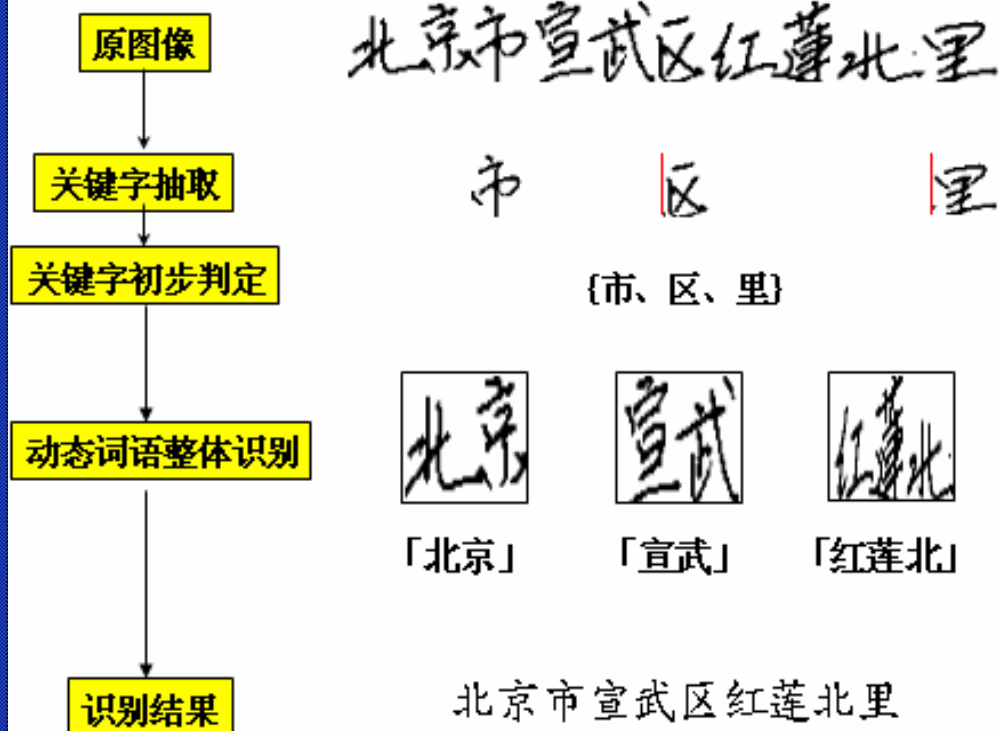
三维漫游交互

汉语手写地址识别

应用目标: 金融保险业票据的手写地址的识别

概要: 基于关键字抽取, 整体识别地址

基于关键字抽取和整体识别的地址识别



样本

上海市上海近郊区浦东南路徽山村

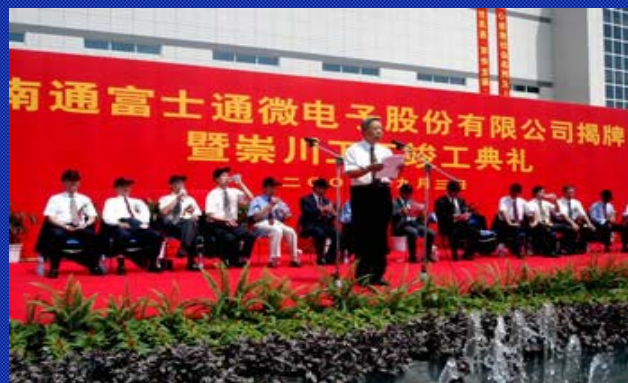
山东省淄博市博山区南家村

文字列识别率: 91.6%
文字识别率: 95.2%

彩色图像中的字符提取和识别

■ 彩色图像中包含众多有价值的文字信息

- 网页图像
- 数码图像
- 杂志图像



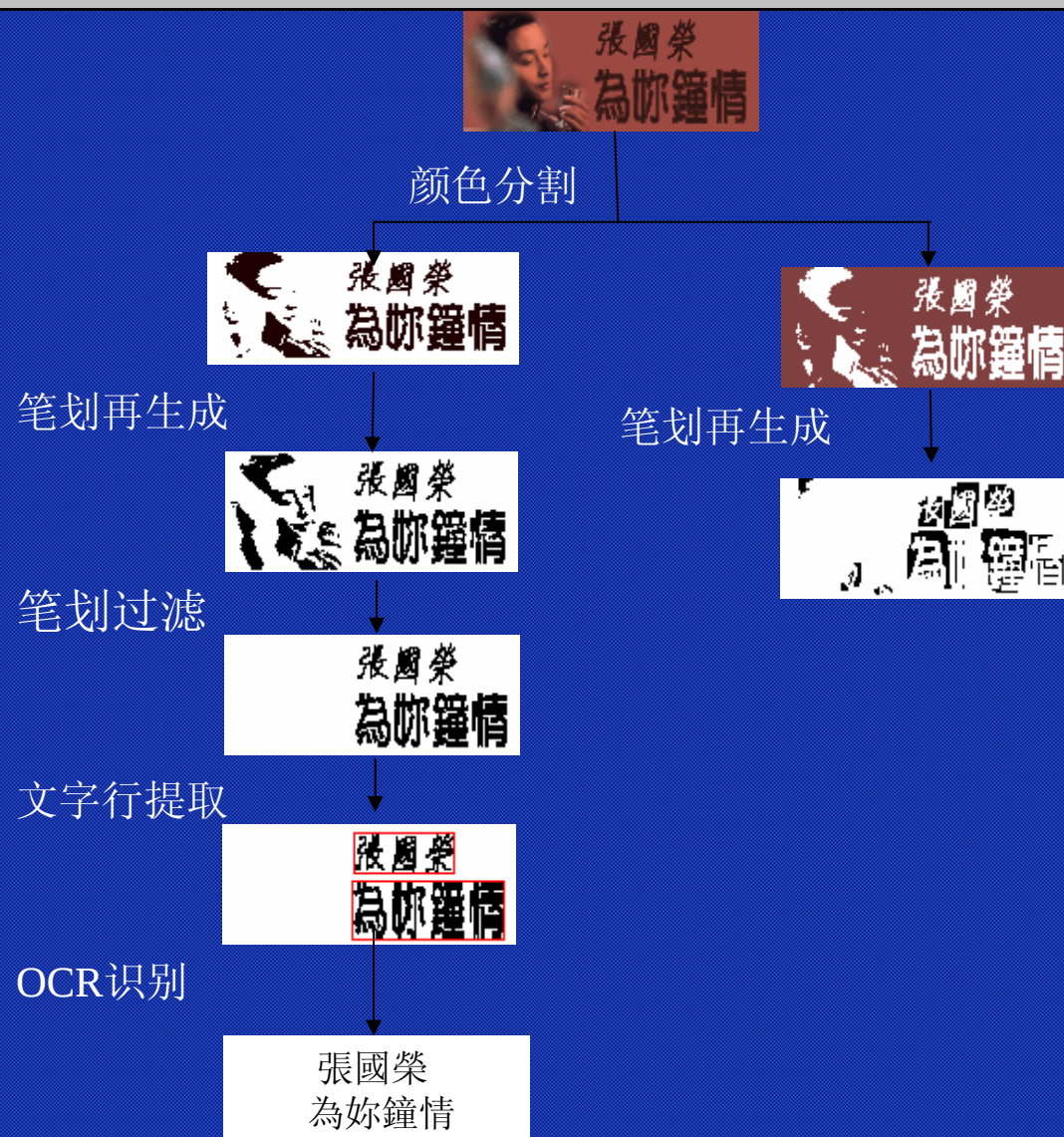
■ 广泛的应用需求

- E-Learning AVI—PPT对应系统
- 互联网内容审查和监视

■ 基于笔画稳定性分析的字符图象提取

基于笔画稳定性分析的字符图象提取

1. 颜色分割
2. 基于稳定性分析的笔划校验
 - 2.1 新笔划生成
 - 2.2 笔划过滤
3. 基于笔划连接的文字行提取
4. 文字识别



电子病历语音识别系统



- ❑ 特定领域的实用化系统
- ❑ 很高的识别准确率和鲁棒性

AmiVoice 超音波診断レポート 音声入力デモ FUJITSU

患者 患者氏名 年齢 性別

脾臓 腎臓 大動脈 その他の所見
肝臓 門脈 胆嚢 肝外胆管 膀胱

辺縁 表面 実質

腫瘍大きさ 形状 エコーレベル 局在

備考 入力開始

病名1 病名2 病名3

Initialized

辺縁 表面 実質

正常

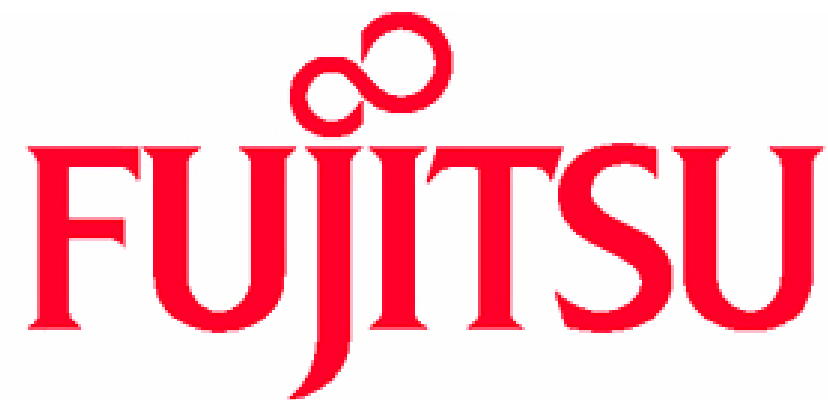
超声波诊断报告 语音输入产品

通过head set mic，凭借声音对
诊断报告的输入可以完成。

对于任意的项目，可以以任意的
顺序输入。

- ❑ 以客户需求为核心进行信息处理
 - 选择可以满足需求的技术
- ❑ 多技术融合，基于内容的信息处理

谢谢！



富士通

网络世界创意无限