

无定义体的表格数据提取技术

【摘要】

随着表格类型的不断增加，银行和保险公司前台的数据输入成本也随之提高。在传统的自动表格提取技术中，即使很多表格的内容相似，但由于格式不尽相同，我们需要预先为每种不同格式的表格定义一个叫做“定义体”的结构以便进行自动识别。为了解决这个问题，我们开发出了“无定义体”的表格识别技术。该技术不是通过表格的格式信息，而是通过表格的逻辑结构来自动提取表格中的数据。在扫描文档（表格，发票等等）以及电子文档（Excel，Word，等）中的数据能够方便地提取出来并用于数据检索和文档内容的提取。本技术通过半自动的输入辅助实现了对数据输入/提取成本的节省。

【技术要点】

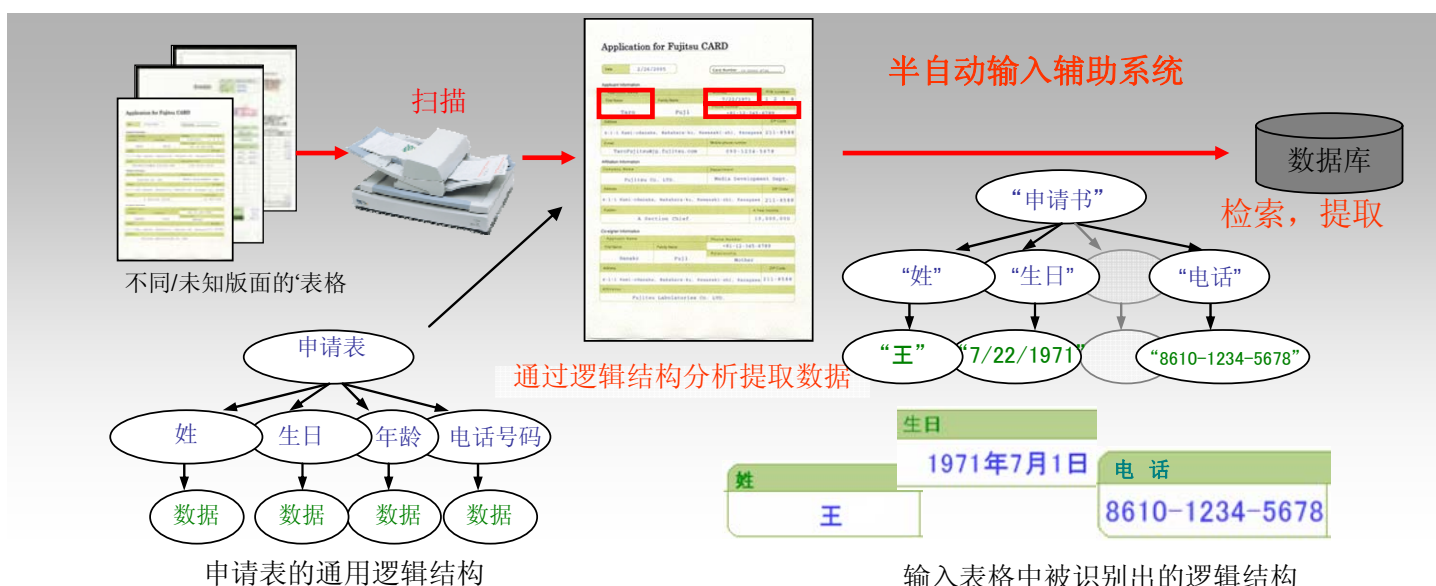
1) 与表格格式无关的“无定义体”表格识别技术

利用表格的逻辑结构，而不是格式信息。首先，根据“通用逻辑结构”提取输入表格的标题（比如，“姓名”，“日期”，等）。在通用逻辑结构中定义了表格中的常用词汇以及建立在词汇意义之上的词汇分级结构。然后，标题对应的数据被自动提取出来。

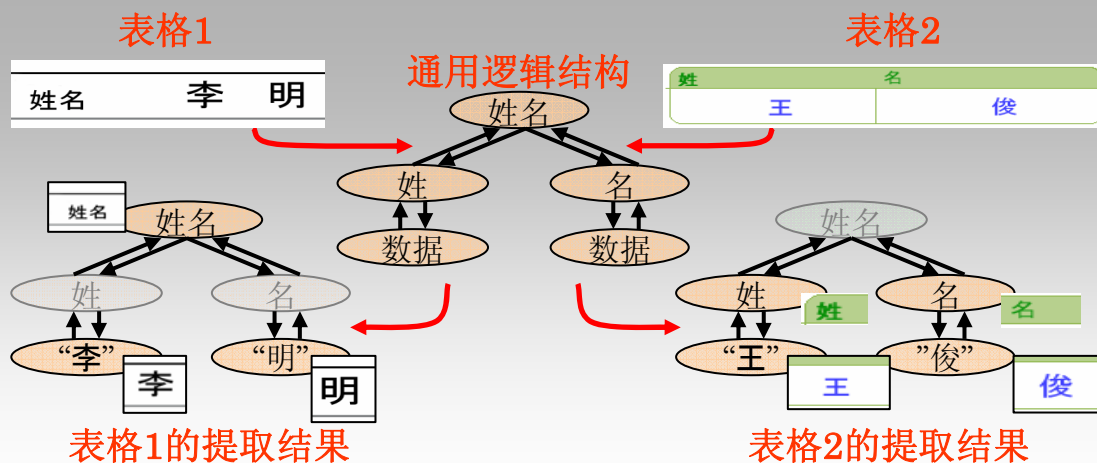
2) 基于概率网络的灵活的逻辑结果分析

我们系统中的概率网络能够解决如下问题：

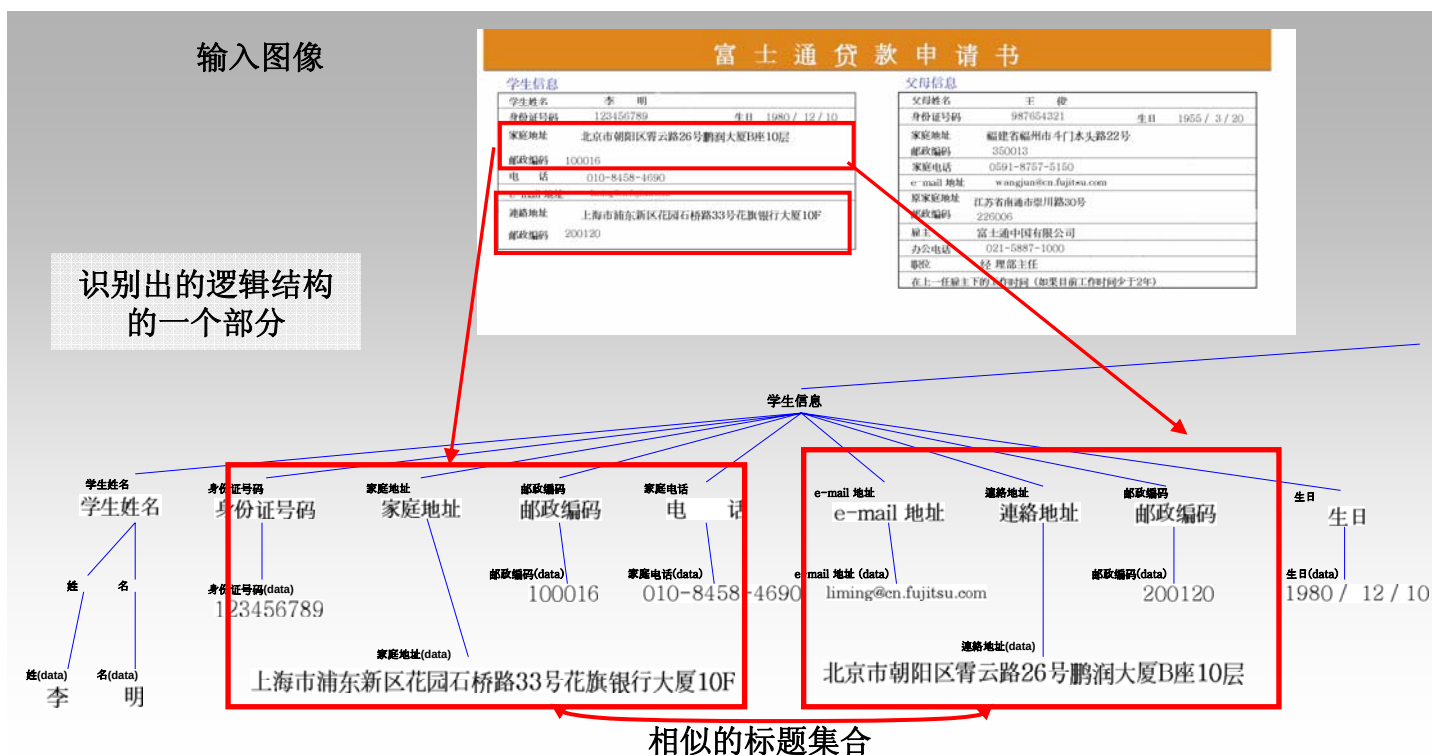
- 标题的变化（比如，“生日”和“出生日期”），
- 由于OCR（光学字符识别）带来的识别错误，
- 逻辑项的缺失。



图一：不同版面格式的表格处理技术



图二：基于“概率网络”的灵活逻辑结构的一个例子



图三：输入图像和分析结果（应用表格）

【应用领域】

1) 前台表格输入辅助系统

本技术通过自动提取需要的数据能够减少各种版面格式的表格输入的成本

2) 用于文档信息提取的数据检索

本技术能够减少在文档信息提取中对扫描表格或者电子文档，如Excel, Word的数据检索成本

3) 多个表格间的合法性一致性检测

提取并比较多个表格内的总额数据以进行公司的内部控制

【联系方式】

孙俊, 信息技术研究部, 富士通研究开发中心有限公司, 北京
e-mail: sunjun@cn.fujitsu.com 010-84584558-529