Network PDF file unstructured information extraction

Pengchanghuan,Sunwei,Fuxiaohan
Shanghai Jiaotong University
598095762@qq.com

Abstract

Modern society is paying more and more attention to the collection and analysis of data.Now we face the annual reports of many companies, we need to intelligently extract and analyze the tables in them.Here comes our project:Network PDF file unstructured information extraction.With our own intelligent analysis extraction program, we have a very high recognition extraction rate (for our test of 32 PDF reports it has a 93% recall rate).

1 Introduction

Modern society is paying more and more attention to the collection and analysis of data.Now we face the annual reports of many companies, we need to intelligently extract and analyze the tables in them. Once successfully acquisition of economic data for most companies, quickly access to market information, the company can make correct decision and successfully occupy a favorable position in the market, so Financial data search engine is of great significance.

The problem we have to face, however, is that the vast majority of financial statements are released in the form of PDF documents.

PDF(Portable Document Format), The purpose of designing PDF file format for Adobe company is to support the information publishing and publishing of Multimedia Integration on cross platform, especially to provide support for network information publication. To this end, PDF has an incomparable advantage over many other electronic document formats. The PDF file format encapsulates text, fonts, formats, colors, and graphics and images independent of device and resolution in a file. The format file can also contain hypertext links, voice and dynamic images and other electronic information, support special documents, integration and security, reliability is higher.

However, the PDF file is a form of documentation designed to prevent modifications (in fact, it is difficult to modify them).So our problem first is to overcome the difficult problem of PDF documents.

We will initially take the text of the PDF file operation, turn it into a CSV file, then the balance sheet to generate a description (how specific description is to design, including fuzzy matching), finally scan the text, positioning the balance sheet, and data extraction.

Our core technical issues are named entity recognition, report description, density based report scanning, positioning, and so on.

As far as our progress is concerned with our own intelligent analysis extraction program, we have a very high recognition extraction rate (for our test of 32 PDF reports it has a 93% recall rate).

2 Our process

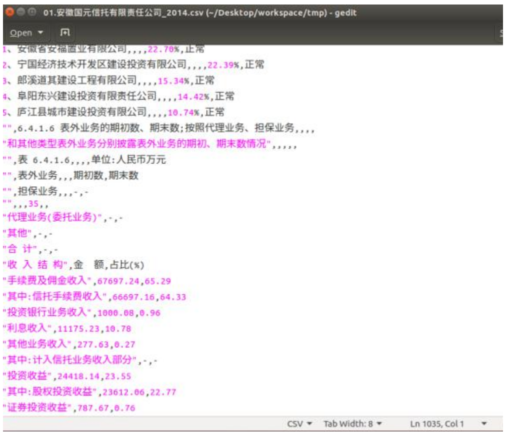
2.1PDF2CSV

Most of the time we get the PDF format.However, there are many limitations to PDF files.For example, you can't copy text, edit data tables, and cannot extract problems like editing Word files.

Tabula[1] can help you extract data tables from PDF files and save them in CSV format so that you can easily access data and use it for the second time.It is an open source free program.

First, we import tabula-java as well as Python library xlwt in our python program.After you enter the correct parameter to these tools,you can transform the entire PDF file into a CSV file.In the CSV file,text and data are separated by commas and line breaks.

Unlike PDF format, CSV is a convenient file type, so we can search, modify and extract the contents of CSV files freely.The Figure 1 shown below is an example of extracted CSV file.


01.安徽国元信托有限责任公司_2014.csv (~/Desktop/workspace/tmp) - gedit
Open
CSV
Tab Width: 8
Ln 1035, Col 1

```
1、安徽省安福置业有限公司,,,,22.70%,正常
2、宁国经济技术开发区建设投资有限公司,,,,22.39%,正常
3、郎溪道其建设工程有限公司,,,,15.34%,正常
4、阜阳东兴建设投资有限责任公司,,,,14.42%,正常
5、庐江县城市建设投资有限公司,,,,10.74%,正常
"",6.4.1.6 表外业务的期初数、期末数;按照代理业务、担保业务,,,,
"和其他类型表外业务分别披露表外业务的期初、期末数情况",,,,,
"",表 6.4.1.6,,,,单位:人民币万元
"",表外业务,,,期初数,期末数
"",担保业务,,,-,-
"",,,35,,
"代理业务(委托业务)",-,-
"其他",-,-
"合 计",-,-
"收 入 结 构",金  额,占比(%)
"手续费及佣金收入",67697.24,65.29
"其中:信托手续费收入",66697.16,64.33
"投资银行业务收入",1000.08,0.96
"利息收入",11175.23,10.78
"其他业务收入",277.63,0.27
"其中:计入信托业务收入部分",-,-
"投资收益",24418.14,23.55
"其中:股权投资收益",23612.06,22.77
"证券投资收益",787.67,0.76
```

Figure 1:Extracted CSV file styles.

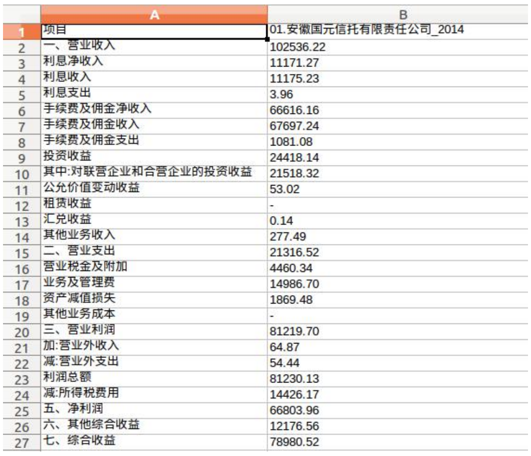
2.2Preliminary extraction

We use the table header specific known as a template for the corresponding part of the matching CSV file, extract the corresponding part of the data, and then stored in our original database, in order to collect the training data we need for fuzzy matching that have the word form and characteristics.

These features include text content, numeric type, numeric size, literal numeric distribution, text density, and so on.

In this experiment, we take 27 known headers of 5 different reports as our initial samples, and all of them are successfully matched to the data in the CSV file.

We can have an example below in Figure 2:



	A	B
1	项目	01.安徽国元信托有限责任公司_2014
2	一、营业收入	102536.22
3	利息净收入	11171.27
4	利息收入	11175.23
5	利息支出	3.96
6	手续费及佣金净收入	66616.16
7	手续费及佣金收入	67697.24
8	手续费及佣金支出	1081.08
9	投资收益	24418.14
10	其中:对联营企业和合营企业的投资收益	21518.32
11	公允价值变动收益	53.02
12	租赁收益	-
13	汇兑收益	0.14
14	其他业务收入	277.49
15	二、营业支出	21316.52
16	营业税金及附加	4460.34
17	业务及管理费	14986.70
18	资产减值损失	1869.48
19	其他业务成本	-
20	三、营业利润	81219.70
21	加:营业外收入	64.87
22	减:营业外支出	54.44
23	利润总额	81230.13
24	减:所得税费用	14426.17
25	五、净利润	66803.96
26	六、其他综合收益	12176.56
27	七、综合收益	78980.52

Figure 2:Extracted tables and data which will be used as raw samples.

2.3 Autodetect module

We obtain a similar approach to intelligent extraction of tabular data features in a single article called:”Research on web core block extraction algorithm based on DOM node text density”[2].Then we develop our own autodetect module.

We find that the appearance of tables is related to many parameters.Such as text content, numeric type, numeric size, literal numeric distribution, text density, and so on.We match the model trained by the existing data in the database with the contents in the CSV file,we compare the feature sets between them.Then select the higher matching part as a table.

This is the processes of our fuzzy

matching for intelligent table extraction.

In addition,We also take the line distribution in the graph of the PDF file as the basis for the table to appear.In this process, we also use tabula as a tools ,and we modify the code to make the tool meet our experimental requirements.

2.4 Visualization

At the same time, we have made some visualization of the extraction of tables to facilitate the use of developers and users.

When using, we can use manual identification methods to find the forms which have not been successfully recognized. All the tables are extracted correctly under manual surveillance

2.5 Convert to CSV file

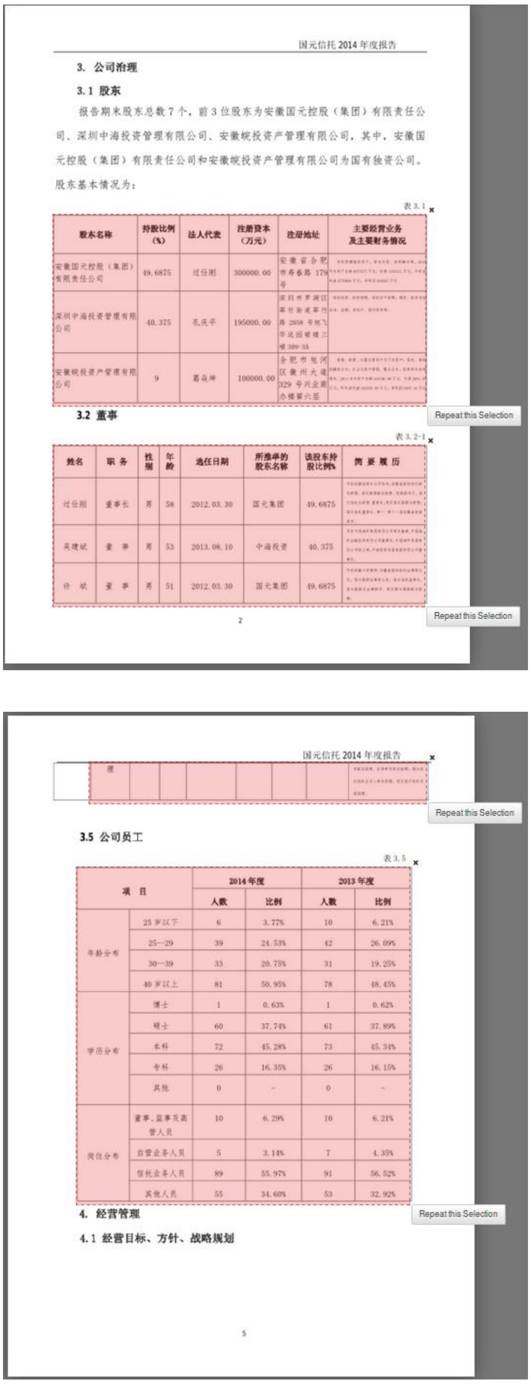
This is the final part of out project,we transfer the tables we extracted into CSV form,in order to make people have a good reading experience.

3 Evaluation

In the test, we used 32 PDF file.We have a recall rate about 93%,such a high recall rate indicates that our program is of practical value.

3.1Table search

In this part we search the tables using fuzzy matching.We will have a preview below in Figure3:



国元信托 2014 年度报告

3. 公司治理

3.1 股东

报告期末股东总数 7 个，前 3 位股东为安徽国元控股（集团）有限责任公司、深圳中海投资管理有限公司、安徽皖投资产管理有限公司，其中，安徽国元控股（集团）有限责任公司和安徽皖投资产管理有限公司为国有独资公司。股东基本情况为：

表 3.1

股东名称	持股比例（%）	法人代表	注册资本（万元）	注册地址	主要经营业务及主要财务情况
安徽国元控股（集团）有限责任公司	49.6875	过仕刚	300000.00	安徽省合肥市寿春路 179 号	[illegible]
深圳中海投资管理有限公司	40.375	孔庆平	195000.00	深圳市罗湖区翠竹街道翠竹路 2058 号旭飞华达园裙楼三楼 309-3A	[illegible]
安徽皖投资产管理有限公司	9	葛焱坤	100000.00	合肥市包河区徽州大道 329 号兴业商办楼第六层	[illegible]

3.2 董事

表 3.2-1

姓名	职 务	性别	年龄	选任日期	所推举的股东名称	该股东持股比例%	简 要 履 历
过仕刚	董事长	男	58	2012.03.30	国元集团	49.6875	[illegible]
吴建斌	董 事	男	53	2013.08.10	中海投资	40.375	[illegible]
许 斌	董 事	男	51	2012.03.30	国元集团	49.6875	[illegible]

2

国元信托 2014 年度报告

理							[illegible]

3.5 公司员工

表 3.5

项 目		2014 年度		2013 年度	
		人数	比例	人数	比例
年龄分布	25 岁以下	6	3.77%	10	6.21%
	25—29	39	24.53%	42	26.09%
	30—39	33	20.75%	31	19.25%
	40 岁以上	81	50.95%	78	48.45%
学历分布	博士	1	0.63%	1	0.62%
	硕士	60	37.74%	61	37.89%
	本科	72	45.28%	73	45.34%
	专科	26	16.35%	26	16.15%
	其他	0	-	0	-
岗位分布	董事、监事及高管人员	10	6.29%	10	6.21%
	自营业务人员	5	3.14%	7	4.35%
	信托业务人员	89	55.97%	91	56.52%
	其他人员	55	34.60%	53	32.92%

4. 经营管理

4.1 经营目标、方针、战略规划

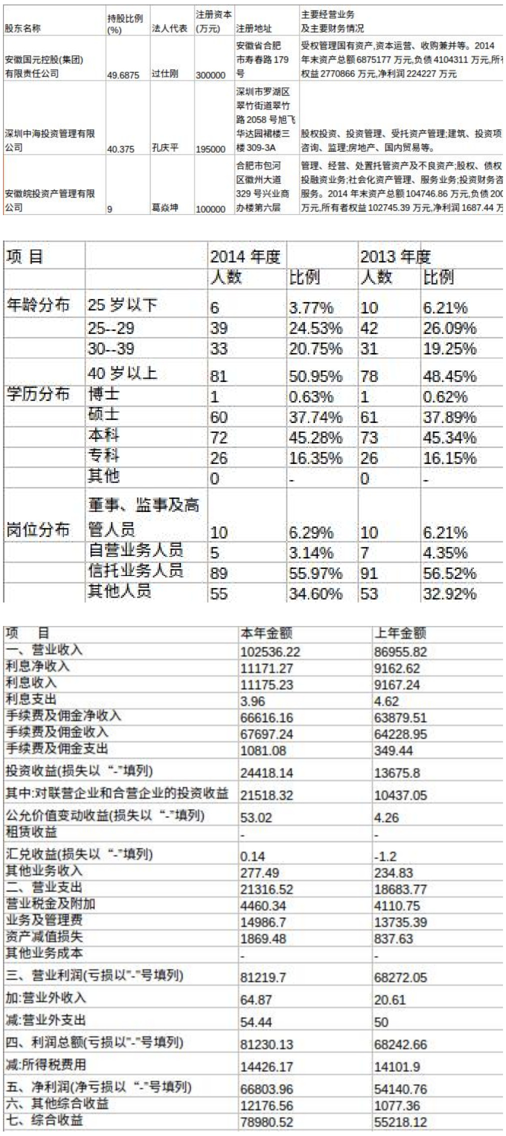
5

Figure 3:Visualization of table search.

In this part,if there are something wrong with the table search,we can manually correct the result by putting a box on the table that is missed.

3.2Table extraction

This is our final part we extract our tables ,transfer them into CSV files.We will have a preview below in Figure 4:



股东名称	持股比例(%)	法人代表	注册资本(万元)	注册地址	主要经营业务及主要财务情况
安徽国元控股(集团)有限责任公司	49.6875	过仕刚	300000	安徽省合肥市寿春路 179 号	受权管理国有资产,资本运营、收购兼并等。2014 年末资产总额 6875177 万元,负债 4104311 万元,所 权益 2770866 万元,净利润 224227 万元
深圳中海投资管理有限公司	40.375	孔庆平	195000	深圳市罗湖区翠竹街道翠竹路 2058 号旭飞华达园裙楼三楼 309-3A	股权投资、投资管理、受托资产管理;建筑、投资项 咨询、监理;房地产、国内贸易等。
安徽皖投资产管理有限公司	9	葛焱坤	100000	合肥市包河区徽州大道 329 号兴业商办楼第六层	管理、经营、处置托管资产及不良资产;股权、债权 投融资业务;社会化资产管理、服务业务;投资财务咨 服务。2014 年末资产总额 104746.86 万元,负债 20 万元,所有者权益 102745.39 万元,净利润 1687.44 万

项 目		2014 年度		2013 年度	
		人数	比例	人数	比例
年龄分布	25 岁以下	6	3.77%	10	6.21%
	25--29	39	24.53%	42	26.09%
	30--39	33	20.75%	31	19.25%
	40 岁以上	81	50.95%	78	48.45%
学历分布	博士	1	0.63%	1	0.62%
	硕士	60	37.74%	61	37.89%
	本科	72	45.28%	73	45.34%
	专科	26	16.35%	26	16.15%
	其他	0	-	0	-
岗位分布	董事、监事及高管人员	10	6.29%	10	6.21%
	自营业务人员	5	3.14%	7	4.35%
	信托业务人员	89	55.97%	91	56.52%
	其他人员	55	34.60%	53	32.92%

项 目	本年金额	上年金额
一、营业收入	102536.22	86955.82
利息净收入	11171.27	9162.62
利息收入	11175.23	9167.24
利息支出	3.96	4.62
手续费及佣金净收入	66616.16	63879.51
手续费及佣金收入	67697.24	64228.95
手续费及佣金支出	1081.08	349.44
投资收益(损失以 "-"填列)	24418.14	13675.8
其中:对联营企业和合营企业的投资收益	21518.32	10437.05
公允价值变动收益(损失以 "-"填列)	53.02	4.26
租赁收益	-	-
汇兑收益(损失以 "-"填列)	0.14	-1.2
其他业务收入	277.49	234.83
二、营业支出	21316.52	18683.77
营业税金及附加	4460.34	4110.75
业务及管理费	14986.7	13735.39
资产减值损失	1869.48	837.63
其他业务成本	-	-
三、营业利润(亏损以"-"号填列)	81219.7	68272.05
加:营业外收入	64.87	20.61
减:营业外支出	54.44	50
四、利润总额(亏损以"-"号填列)	81230.13	68242.66
减:所得税费用	14426.17	14101.9
五、净利润(净亏损以 "-"号填列)	66803.96	54140.76
六、其他综合收益	12176.56	1077.36
七、综合收益	78980.52	55218.12

Figure 4:Extracted tables.

4 Discussion

In our project we can't extract the title of the table,because it can exist in all directions of the table.So the next thing to do is to look for the title of the table intelligently.

In this project,the senior gave me the train of thought,then, the following program editing, data collection, tool search, and, most importantly, the results are all done by myself. This article does not have much theoretical knowledge, this project uses practical as the goal, but also hope that the teacher can give more guidance to me.

5 Conclusion

With our own intelligent analysis extraction program, we have a very high recognition extraction rate (for our test of 32 report draws have 93% recall rate).

Reference

[1] http://tabula.technology/

[2]http://xueshu.baidu.com/s?wd=paperuri%3A%284cc2a28db2c77a0e3a929d54ef9ca624%29&filter=sc_long_sign&tn=SE_xueshusource_2kduw22v&sc_vurl=http%3A%2F%2Fcdmd.cnki.com.cn%2FArticle%2FCDMD-10007-1012007246.htm&ie=utf-8&sc_us=18388595119318769597