Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Data service generation framework from heterogeneous printed forms using semantic link discovery

Han Yu, Hongming Cai [*], Jun Zhou, Lihong Jiang

School of Software, Shanghai Jiao Tong University, 800 Dongchuan Rd, Shanghai, China

HIGHLIGHTS

- A complete and feasible framework from printed forms to data service is proposed.
- An automatic form data extraction and structuration approach is presented.
- A usable prototype system integrating heterogeneous printed resumes is implemented.

ARTICLE INFO

Article history:
Received 28 March 2017
Received in revised form 7 September 2017
Accepted 22 September 2017
Available online xxxx

Keywords:
Form recognition
Heterogeneous data integration
Semantic data model
Table matching
Data service generation

ABSTRACT

Printed forms contain rich information in business process and daily life. However, tremendous heterogeneous printed forms containing same categories of information are difficult to manage and share, which lead to massive data in printed forms remaining waste. To automatically integrate and share these data remarkably improves the efficiency of enterprises, the key problem is how to extract heterogeneous data in printed forms and integrate them for quick use. To solve this issue, we propose a framework that discovers semantic links in printed forms and generates data services for easy data management and rapid data sharing in the enterprise systems. First, a multiple-OCR-based form recognition approach is proposed to make forms computer-readable. Next, forms are modeled into semi-structured data using structure-based semantic link discovery and refining with massive data. Then, a linked data model is built by table matching to align data. Finally, data services are generated based on the linked data model. A series of experiments on printed resumes are conducted, and the results illustrate our framework performs well in recognition rate, link discovery accuracy, data compression ratio and data resource accuracy. A prototype system is presented to illustrate the feasibility of the proposed framework.

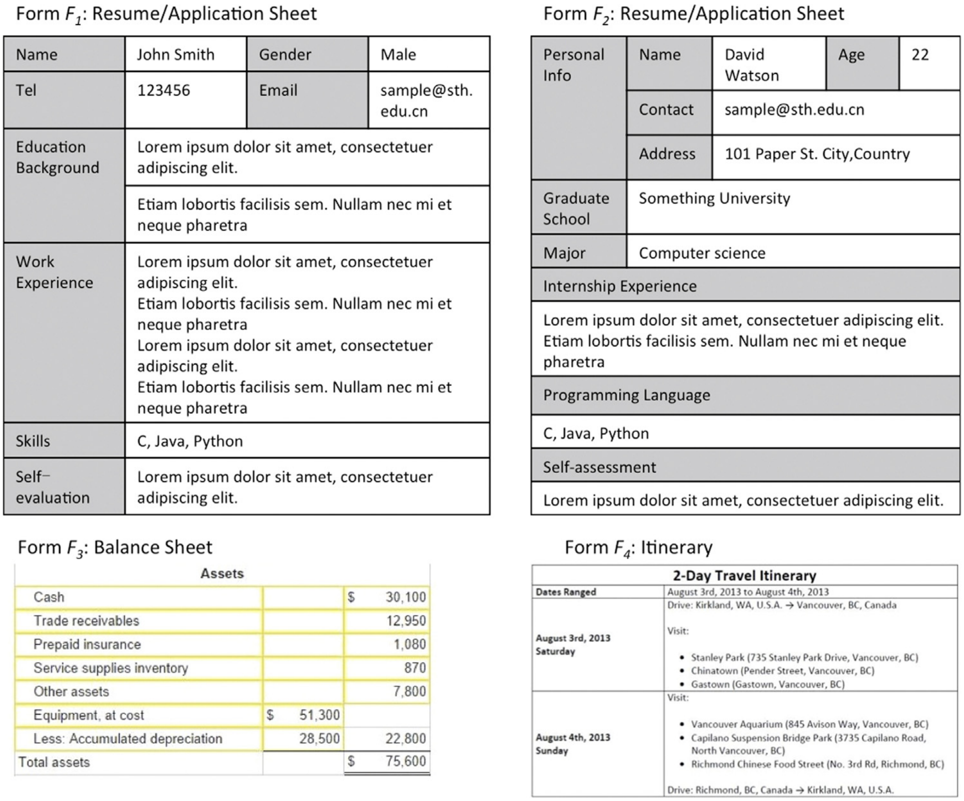
1. Introduction

The forms are one of the most commonly used data carriers in business and daily life: HRs need to select qualified candidates for the company by looking through hundreds of various resumes a day, auditors need to go through hundreds of heterogeneous balance sheets to find mistakes, travel agents provide different kinds of itineraries for different travelers. There are still millions of printed forms, even though digitalization revolution has been developed for tens of years. The rich information in the printed forms is still useful and irreplaceable. In the era of the Internet, information is power. Better data integration and sharing mean more efficient work and further bring more benefits. Therefore, the rapid and convenient information retrieving, communication and management become the most significant processes valued by the enterprises. Thus, digitalizing printed forms and extracting rich information for data integration and sharing will greatly benefit the enterprises by improving the efficiency of data communication and management.

However, with no unified standard or template, people turn to use heterogeneous forms to describe instances in same categories, such as resumes(F_1, F_2 in Fig. 1), balance sheets (F_3 in Fig. 1), itineraries (F_3 in Fig. 1). These forms have several notable features. To that end, (1) **They are heterogeneous in both structure and semantics.** Take F_1 and F_2 in Fig. 1 as an instance, these two forms are both resumes but have different organizing structures and diverse property fields(gray-background cells in form). Furthermore, among these fields some have diverse expressions but semantically represent the same concept, like `Self-evaluation` in F_1 and `Self-assessment` in F_2. (2) **Forms are loosely connected.** Compared to long text, there are no sentence components, syntax, and grammars in forms but only form lines and discrete expressions with almost no connection words like “is”, “for”, etc.,

[*] Corresponding author.
E-mail addresses: sharonhanz@sjtu.edu.cn (H. Yu), hmcai@sjtu.edu.cn (H. Cai), zj1129@sjtu.edu.cn (J. Zhou), jiang-lh@cs.sjtu.edu.cn (L. Jiang).

https://doi.org/10.1016/j.future.2017.09.059
0167-739X/© 2017 Elsevier B.V. All rights reserved.



Form F_1: Resume/Application Sheet

Name	John Smith	Gender	Male
Tel	123456	Email	sample@sth.edu.cn
Education Background	Lorem ipsum dolor sit amet, consectetuer adipiscing elit.		
	Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra		
Work Experience	Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra		
Skills	C, Java, Python		
Self-evaluation	Lorem ipsum dolor sit amet, consectetuer adipiscing elit.		

Form F_2: Resume/Application Sheet

Personal Info	Name	David Watson	Age	22
	Contact	sample@sth.edu.cn		
	Address	101 Paper St. City,Country		
Graduate School	Something University			
Major	Computer science			
Internship Experience				
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra				
Programming Language				
C, Java, Python				
Self-assessment				
Lorem ipsum dolor sit amet, consectetuer adipiscing elit.				

Form F_3: Balance Sheet

Assets		
Cash		$ 30,100
Trade receivables		12,950
Prepaid insurance		1,080
Service supplies inventory		870
Other assets		7,800
Equipment, at cost	$ 51,300	
Less: Accumulated depreciation	28,500	22,800
Total assets		$ 75,600

Form F_4: Itinerary

2-Day Travel Itinerary	
Dates Ranged	August 3rd, 2013 to August 4th, 2013
August 3rd, 2013 Saturday	Drive: Kirkland, WA, U.S.A. → Vancouver, BC, Canada Visit: • Stanley Park (735 Stanley Park Drive, Vancouver, BC) • Chinatown (Pender Street, Vancouver, BC) • Gastown (Gastown, Vancouver, BC)
August 4th, 2013 Sunday	Visit: • Vancouver Aquarium (845 Avison Way, Vancouver, BC) • Capilano Suspension Bridge Park (3735 Capilano Road, North Vancouver, BC) • Richmond Chinese Food Street (No. 3rd Rd, Richmond, BC) Drive: Richmond, BC, Canada → Kirkland, WA, U.S.A.

Fig. 1. Common printed form examples.

which leads to the third feature that (3) **forms are not computer-understandable.** Due to these features, there are no automatic approaches to deal with data in printed forms. Meanwhile, manually reading and processing these forms cost a large amount of time and man power to cope with these forms and utilize data from these forms for further use. The core problem we are faced with is how to extract and integrate heterogeneous data from printed forms.

The problem can be resolved into four key problems to solve: (1) We need to recognize forms and transfer them into computer files, which is an OCR issue. There already exist some sophisticated OCR techniques, such as Tesseract [1] that allows data training to obtain better performance in particular cases, and ABBYY[1] that widely supports rich-text documents. However, these OCR engines perform unsatisfyingly in recognizing form format documents. (2) We want the computers to automatically understand expressions in forms and semantic relations among them, which is a semantic link discovery problem. Nowadays, much semantics research has been done on linked data, for examples, key discovery [2], association discovery [3] and long-text analysis based on open data [4] and context [5]. However, the solution to semantic discovery on forms has not been proposed yet. (3) How to merge heterogeneous forms with various fields into a single linked data model, which is similar to the table-matching problem. Some [6] matches short tables to open data, some [7] matches data from different open data sources, and some [8] matches terms incrementally. Our matching target is large and rich forms, which is different from above approaches. However, the matching methods can be learned from former research. (4) How to provide data services based on the linked data model, which is a data service generation problem. R.T. Fielding's RESTful [9] is a widely used and lightweight web service architecture. Based on it, some sophisticated approaches from ontology [10] and general data model [11] to data service are proposed and applied in practice.

In this paper, we propose a complete framework for generating data services from heterogeneous printed forms. First, we combine two OCR techniques for the form recognition. Then, we propose a new semantic link discovery approach to model the forms automatically. Next, we use table matching approach and build a linked data model to store all form instances. Finally, we provide data service based on the linked data model, including basic data services like retrieving and search, plus special services like data homogenization and evaluation. Furthermore, we conduct a series of experiments on our framework and build a prototype system based on the framework to reflect the feasibility and usability. The contributions of this paper can be summarized as follows:

- A feasible framework is designed for first to generate data services from heterogeneous printed forms. It is a complete process from printed forms directly to data services that has not been proposed before.
- An adoptable data extraction method is presented. First, an improved form recognition approach is proposed. Then a two-phase semantic link discovery method is applied to generate relation model. Both methods perform acceptably in experiments and are usable in practice.
- Data integration and service generation are first applied in printed forms and a prototype system is implemented illustrating the usage of the framework.

The paper is organized as follows: First, Section 2 introduces the framework of our approach, and Section 3 illustrates the details of methods used in the framework. Then, Section 4 shows how we conduct experiments and the results. Section 5 presents a prototype system based on our framework. Finally, Section 6 introduces related works and Section 7 concludes the whole paper.

[1] http://ocrsdk.com/.