COMPARING THE LINE CROSSING FEATURE APPROACH WITH THE GRAPH APPROACH FOR INVOICE AUTOMATION

Lydia Mutiara Dewi
Fakultas Ekonomi, Universitas Katolik Parahyangan

Abstrak

Di dalam dunia bisnis, invoice merupakan salah satu dokumen penting yang erat kaitannya dengan aktivitas penjualan dan pembelian di dalam suatu perusahaan. Setiap invoice yang diperoleh oleh suatu perusahaan akan menjadi sumber informasi atas berapa besarnya hutang atau piutang yang dimiliki oleh suatu perusahaan. Ketika jumlah transaksi di dalam suatu perusahaan masih tidak terlalu banyak, pencatatan invoice ke dalam database perusahaan secara manual masih dimungkinkan. Mengingat bahwa jumlah transaksi di dalam suatu perusahaan dapat menjadi besar, suatu metode untuk melakukan pencatatan invoice secara otomatis dapat membantu proses pencatatan data ke dalam database perusahaan agar menjadi lebih efisien. Dua pendekatan pencatatan invoice secara otomatis yang akan dibandingkan di dalam karya ilmiah ini adalah Line Crossing Feature Approach dan Graph Approach. Berdasarkan perbandingan atas kelebihan dan kekurangan masing-masing pendekatan, Graph Approach secara pribadi dapat dinilai sebagai sebagai pendekatan yang lebih tepat untuk pencatatan invoice secara otomatis karena fleksibilitasnya dalam mengenali berbagai jenis dokumen.

Kata kunci*: invoice, information extraction, Line Crossing Feature Approach, Graph Approach, INFORMys method*

1. Introduction

In the business world, companies, especially those that are engaged with the sale of goods, are inseparable from the functions of sales and purchasing. Principally, a company buys goods for later resale either to distributors or to end consumers in order to get the optimum benefit. Of course when a company buys goods, a company has to pay. A company has choices either to pay in full amount or to settle the payment in several stages. Most of the companies prefer to choose the latter than the first. The reason is that they can use the rest of the money for optimizing their business performance through activities that can increase the cash inflow of a company.

If a company chooses for the second option, then a company should have a proper record about how much the amount of money they owed or owing. An invoice is a source of document that supports the report of Account Payable and Account Receivable. These records show the total amount of money which is owed or owing by the company at the end of the period.

According to West's Encyclopedia of American Law, The term “invoice”[1] refers to a written account or itemized document which contains the list of goods send to the buyer or consignee by the vendor that indicates the information for each piece of merchandise, such as the quantity and price. This document indicates that the buyer has to pay the seller in the maximum number of days. In the seller point of view, this document is called sales invoice whether in the buyer point of view this is called purchase invoice.

In the past time, a company proceeded to record every single invoice manually. When the scale of a company is getting big and the amount of transactions is getting huge, the manual input is no longer efficient. Starting from this point of view, company starts to develop a technology called Invoice automation that could record the information on every single invoice automatically by using basic approaches of Information Retrieval. I restrict the scope of invoice to the term of “purchase invoice” in this paper.

2. Information Extraction

Suppose a company received a vast amount of invoices from the seller, and a company wants to have an automatic input for the purpose of making account payable report, a company needs to scan all the documents by means of invoices. After the documents are scanned, then the Optical Character Recognition (OCR) technology is implemented. According to Encyclopedia Brittanica Online, OCR[2] is the technique to scan and to compare with the purpose of identify the printed text or numerical data. This technoloy allows us not to retype the printed version data for the purpose of data entry.

For the simplicity reason, let us assume that the OCR technology works perfectly. Even though the OCR technology works out, the other question according to the data entry occurs. How can computer recognize the form of the document and its contents, and extract the useful information from the document like humans do? To answer this question, one needs a specific technology, namely Information Extraction. Information Extraction (Kaiser & Miksch, 2005) is one of the Natural Language Processing Technology which has the purpose of processing unstructured, natural language text to detect the specific information pieces or facts in the text, and use these facst to fill in a database. In other words, information extraction is an effective way to populate the contents of relational database.

According to Jurafsky and Martin (2009), there are four steps of doing Information extraction:

[1] See http://www.encyclopedia.com/topic/Invoice.aspx#1

[2] See http://www.britannica.com/EBchecked/topic/430371/OCR

- Named entitiy recognition, which is to detect and classify all the proper names that appear in text. In this step, one meets the term of named entity mention that contains some instances with a proper name like organization, people, place, times or amounts (for instance: seller name, buyer name, amount that has to be paid, and so forth). In this step one also needs a reference solution to check whether the entity that appears in a certain location (e.g a line in a document) is the same entity with the one which appears in another location (e.g. another line in a document). The reference solution is usually used in a document that contains several lines and using the entity name repetitively.
- Relation detection and classification is to detect and classify semantic relations among the entities which appear on the text. For instance: “Lydia is an employee of the purchasing department at Company A”.
- Event detection and classification is to indicate and classify in which event the entity is participating, for instance: “Lydia is doing a purchase”.
- Temporal expression recognition and temporal analysis is correlated to detect when the event in text occurs and how they relate to each other. The temporal expression detection tell us that our sample text contains temporal expressions such as day, week, month, etc and relative expressions contains some phrases like two days from now, next month, and so forth. While the temporal expression detects if the sample contains several expressions, the temporal analysis is being used to map the temporal expressions onto the specific calendar (date or times of day) and used that information to situated events in time.

Based on the general overview above, I will discuss two approaches whose functions are to recognize documents (specifically purchase invoice in this case) for the purpose of achieving the goal of information extraction. The first approach is identifying document with the Line Crossing Signature Approach and the second one is extracting information using the Graph Approach (INFORMys technology).

3. The Line Crossing Signature Approach

This first approach is to concentrate on processing the preprinted forms with the printed horizontal and vertical lines as the field delimiters. The approach suggested by Taylor, Fritzon, and Pastor (1992) is demonstrated on the United States Internal Revenue Service forms (e.g. tax form). The general steps of this approach which as illustrated in the Figure 1 is as follow:

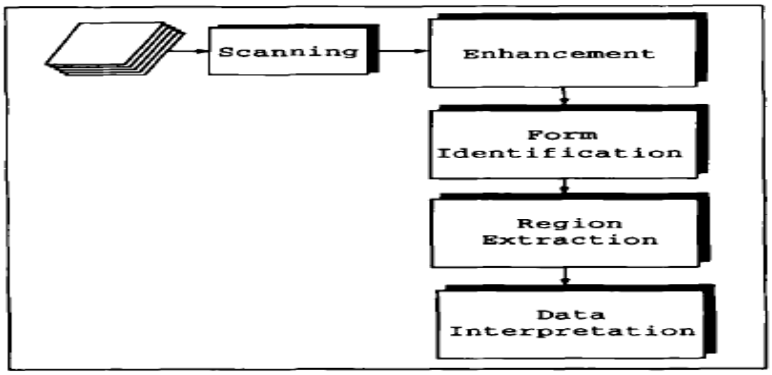

Figure 1: Image Process for Form Analysis
(Extraction of Data from Preprinted Forms, 1992, p. 212)

Description:

1 Scanning : the process to transform the paper form into the electronic form

2 Enhancement : the step of processing image so it can be used for the next operations, for instances: noise removal, tresholding, and skew corrections.

3 Form Identification : the stage to determine particular type of form

4 Region Extraction : the step to locate and identify the important fields of information on the form.

5 Data Interpretation : the step to convert the extracted regions from the image representation to the text representation such as ASCII

From all of the steps above, Taylor et al. (1992) also gave further explanation starting from the enhancement process to the region extraction process. The explanation with the illustration on the figure 2 is as follow:

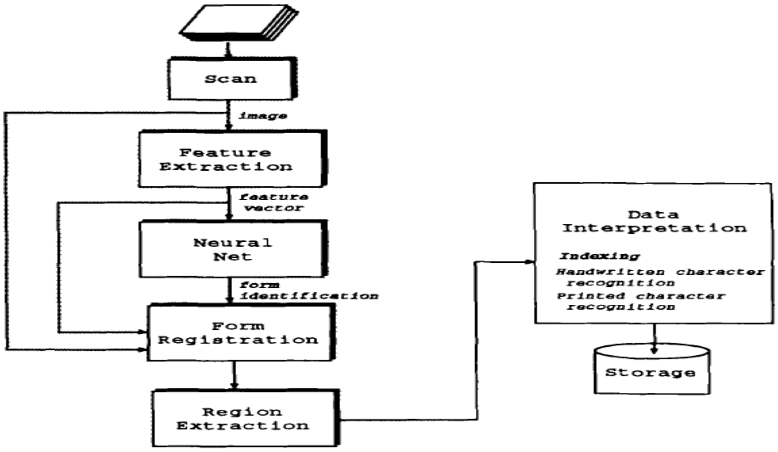

Figure 2: Feature Extraction Analysis
(Extraction of Data from Preprinted Forms, 1992, p. 213)

Description:

- Document is scanned and transformed to the digital image.The digital image is transformed to the feature vector representation based on the line crossing feature. The illustration of the possible line crossing is discribed in the the following Figure 3:

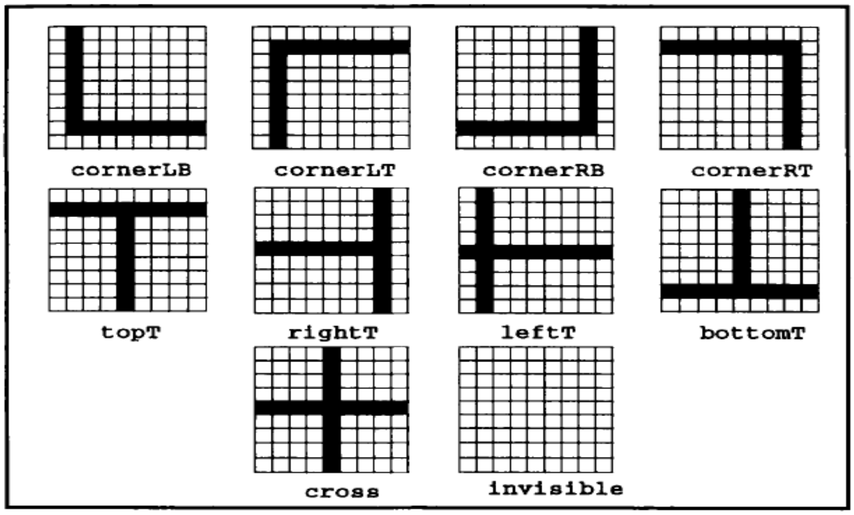

Figure 3: Illustration of possible line crossing
(Extraction of Data from Preprinted Forms, 1992, p. 215)

The invisible part is an imaginary corner that is needed to complete the description of the field but it's not visible on the form. The illustration of the application for line crossing is indicated in the following Figure 4:

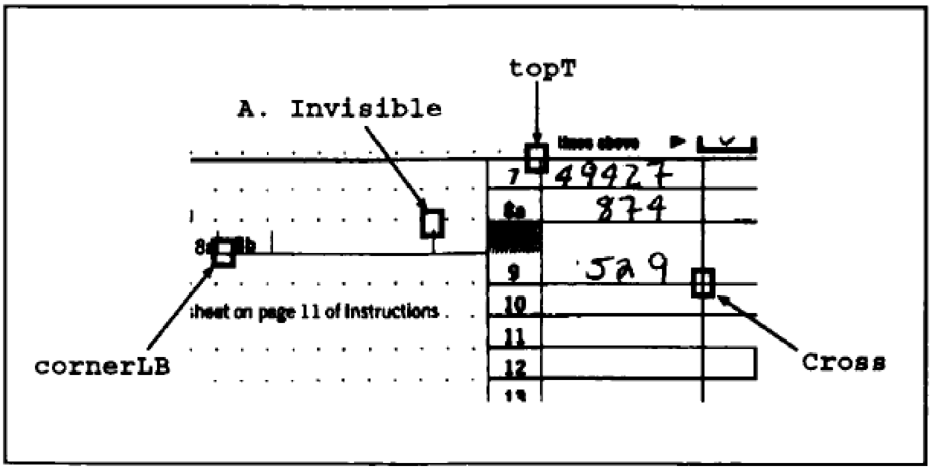

Figure 4 : Illustration of the application for line crossing
(Extraction of Data from Preprinted Forms, 1992, p. 215)

- The feature vector representation, as it is indicated on the Figure 4 is being used as an input for (artifical) neural network and image registration to its model after the specific form is determined. Gerhenson (2001) describes Artificial Neural Network as a computational model which is inspired by the biological neuron. The neural network is well suited for resolving the problem of proposing the form model since we can use the line crossing feature to reduce representation of the data, moreover we have to distinguish one forms from the others. The neural network approach used in this case implements the scheme of 1-of-N input of the coding of the neural net. Using this approach we can rank the identification of the result on the form. If the identification of the proposed form model fails, we can use another form.
- After the input is registered, the important field is checked for the information content. If the content is not empty, it will be extracted.
- The next step is indexing the documents. In this Internal Revenue Service case, the indexing process is implemented by means of replacing the pre-printed label on the form, of correcting the label for skew, and of reading the label supported by OCR software. If there is no label, then the typed information or handwriting is extracted during the field location and extraction stage.
- The region itself can be used as an input for printed or handwritten Character recognition. Below is a figure of the completed IRS form with extracted region:

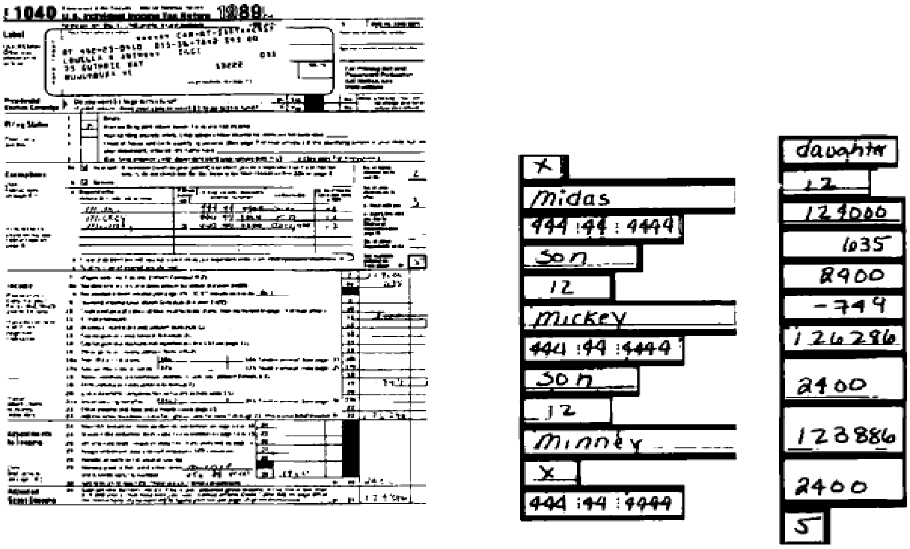
1040
1989
X
midas
444:44:4444
son
12
mickey
444:44:4444
son
12
minney
X
444:44:4444
daughtr
12
124000
1035
8900
-749
126286
2400
123886
2400
5

Figure 5: Sample IRS document with extracted field
(Extraction of Data from Preprinted Forms, 1992, p. 221)

- The extracted information is stored in the database.

4. The Graph Approach

The other method for doing invoice automation which I mentioned before is the Graph Approach. One type of method in the graph approach which I consider for this paper is the INFORMys technology. As it is described by Cesarini, Gori, Marinai, and Soda (1998), the INFORMys (stands for “flexible **IN**voice-like **FORM**-reader s**ys**tem”) is a technology which is specifically arranged to handle down the problem of known-class forms as a common matter in an accounting company. The INFORMys approach is based on a graph whose function is to describe a form of layout. The type of graph used in this approach implements a non hierarchical Attribute Relational Graphs, or abbreviated as ARG.

ARG is commonly used in the computer vision field. This approach is using two components that are nodes and arcs. Nodes describe objects or parts of the objects, while arcs describe mutual relationship between nodes by the meaning of numerical attribute.

The nodes represent line, instruction field, and information field and logos, while arcs represent connections of the nodes. The illustration of each components and the explanation of the process describe in Figure 6 below:

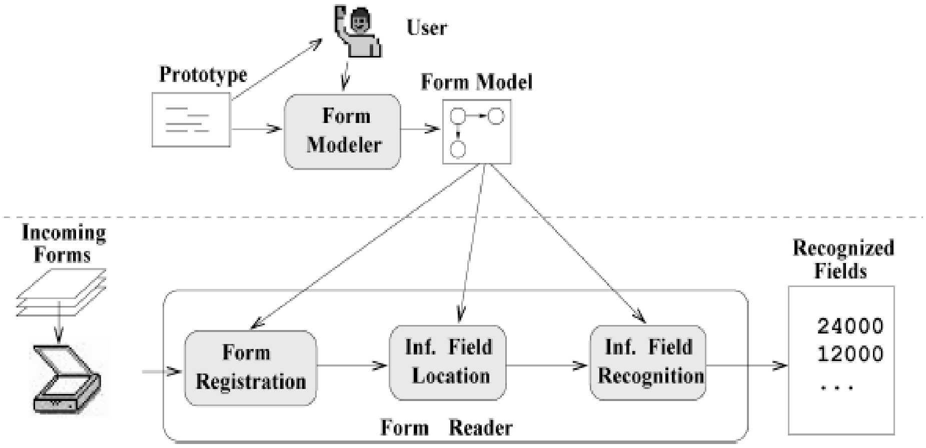

Figure 6: The Basic Structure of INFORMys
(INFORMys: A Flexible Invoice-Like Form-Reader System, 1998, p. 3)

The stages of the INFORMys approach as it is described on Figure 6 by Cesarini et al. (1998) is as follow:

- Form registration

In this initial stage, features that are used to evaluate the form position are stored as nodes in sub-graph and form graph, which is also called with registration graph.

The hypothesized transformation is computed by an aligning model feature with the corresponding feature on the incoming form. The verification of the hypothesized transformation itself is confirmed by searching a form for the items that match with nodes of registered graph. If the verification fails, a new hypothesized will be generated and the hypotheses will be verified again until the verification succeeded or until there are no other hypotheses to be verified.

- Information Field Location (Layout Analysis)

The second stage continues with the layout analysis. This stage focuses on the location of logos and the instruction field, whose function is to describe the content of information field. The instruction field contains some words and it is the way to identify the words inside is by using the whole word as an input of classifier.

- Information Field Recognition

This third stage is also called as the connectionist-based model. In this stage keyword and logo are being considered as an input pattern. This method is useful to deal with the highly noisy form and speed up the recognition process. In this stage information is extracted by segmenting the words using Run Length Smoothing Algorithm (RLSA). Run Length Smoothing Algorithm is a method that can be used to block segmentation and text discrimination (Wong , Casey & Wahl, 1982) .

INFORMys also offered a twofold user interface in order to assist a user during composing a modeling and reading. These twofold user interface forms are:

- The form modeler which is useful to assist user to build the form graph
- The form reader which is useful to implement the recognition engine.

After the components of INFORMys are described, the next stage continues with how this technology actually works. According to Cesarini et al. (1998), below are the works of the graph approach implementing the INFORMys Technology:

- To construct a Form graph. The operator points to the objects, which can be described as line, instruction and information field, and logos that are relevant to represent the structure of the form. Object is corresponds to nodes and arcs express the mutual relationship of different objects.
- To locate the object and their mutual relationship are carried out by user and the corresponding information is exploited to create the form graph.
- To use the form reader module, which support the recognition stage.
- To continue with the form reading which covers the sequential process since the incoming form is scanned, converted to image, aligned by the form register until the information fields are located and stored.

5. Advantages and Disadvantages of Both Methods

These approaches must have their respective advantages and disadvantages. As a conclucion of the descriptions above, I will provide the advantages and the disadvantages of both methods in following table (Cesarini et al., 1988; Taylor et al., 1992):

Description	**Line Crossing Signature Approach**	**Graph Approach**
Advantages	It guarantees that the locations of all data fields in general are identified on the form image even if these data can't be defined by implementing the visible marking on the page.	• The flexibility of processing data which are slightly different from one type of document to another. • Efficiency in the whole system architecture.
Disadvantages	Inability to predict where is the exact location of data field the form when the image is formed on the scanner.	• It might not be the best approach if the structure of document is already known. • The approach will cause a very expensive algorithm.

Table 1: Comparison of advantages and disadvantages of both methods

6. Conclusion

According to both approaches that have been explained above, it can be followed that:

- The Line Crossing Signature approach identifies a form of purchasing invoice by means of scanning and transforming the document into a digital image. The digital image then changed into the feature vector representation in accordance with the line crossing feature. After the form of purchasing invoice is registered, the important fields on the invoice form are checked whether the information is contained or not. If one indicates content inside the field, the content will be extracted.
- Meanwhile the INFORMys approach identifies a form of purchasing invoice by means of an operator whose function is to search the relevant object represented in the structure of the form in accordance with the operator points. After the structure is recognized, the information is extracted by means of segmenting the words using the Run Length Smoothing Algorithm (RLSA).

Looking at the comparison of both methods for invoice automation, I prefer the second approach than the first since the Graph Approach is designed mainly for the accounting documents like purchasing invoice (Cesarini et al.,1998). Moreover, the INFORMys method is flexible enough since it could recognize any types of invoice document. Meanwhile, the Line Crossing Signature approach might have difficulties in predicting the exact location of data fields on an invoice form when the image is formed on the scanner.

References

Cesarini, F., Gori, M., Marinai, S.,& Soda, G.(1998). INFORMys: A Flexible Invoice-Like Form-Reader System. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20, pp. 730-745.

Gershenson, C. (2001). Artificial Neural Networks for Beginners**.** Formal Computational Skills Teaching Package, COGS, University of Sussex. Retreived December 08, 2011 from http://uk.arxiv.org/ftp/cs/papers/0308/0308031.pdf

Invoice. (2005). West's Encyclopedia of American Law. Retrieved December 08, 2011 from Encyclopedia.com: http://www.encyclopedia.com/doc/1G2-3437702397.html

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing, *Information Extraction* (pp. 759-798). Upper Saddle River, New Jersey: Pearson Education, Inc.

Kaiser, K., & Miksch, S. (2005). Information Extraction A Survey. *Technology*, 32, pp. 1-24, Retrieved December 08, 2011 from http://ieg.ifs.tuwien.ac.at/techreports/Asgaard-TR-2005-6.pdf

OCR. (2011). In *Encyclopædia Britannica*. Retrieved December 08, 2011 from http://www.britannica.com/EBchecked/topic/430371/OCR

Taylor, S. L., Fritzson, R., & Pastor, J. A. (1992). Extraction of Data from Preprinted Forms. *Machine Vision and Applications, 5,* pp. 211-222, Retrieved December 13, 2011 from http://www.springerlink.com/content/t173q68gk52rr624/

Wong, K. Y., Casey, R. G., & Wahl, F. M. (1982). Document Analysis System*. IBM Journal of Research and Development,* 26, pp. 647-656.