Information Extraction from Receipts using Machine Learning

Mario Karlovčec[1]
[1]*Jožef Stefan International Postgraduate School, Jamova 39, Ljubljana, Slovenia*
E-mail: mario.karlovcec@ijs.si

Abstract. *Automated information extraction from receipts can help us to easier organize our expenses. Approach to using Machine learning algorithms for information extraction is introduced. Datasets are generated using the developed application which enables labeling of textual documents. Results of applying the Machine learning algorithms on the datasets show usefulness of the approach.*

Keywords. information extraction, machine learning, classification, dataset, labeling.

1. Introduction

The receipts received after purchasing products can be used for managing daily, monthly or yearly spending of a person. This can be done by manually processing the receipts, rewriting its main elements to a piece of paper or into some software that enables various mathematical and statistical analyses. Much better approach would be to process the receipts automatically, using computer algorithms to extract its main elements. In contrast to manual processing, automated approach takes much less of person's time and it is much easier.

Information Extraction is the name given to any process which selectively structures and combines data which is found. [1] We proposed an automated information extraction approach from receipts, based on some machine learning algorithms. To use these algorithms, feature set for the receipt elements is designed. Application which enables labeling of receipt elements and extracting the features from the receipts is created. Experimental results show the performance of proposed approach.

2. Proposed approach

The approach is divided into three main parts: (1) pre-processing of receipt to get a textual file; (2) generating records with defined feature set; and (3) using machine learning algorithms. Proposed approach is shown with figure 1. Each main part is represented in one row of the figure.

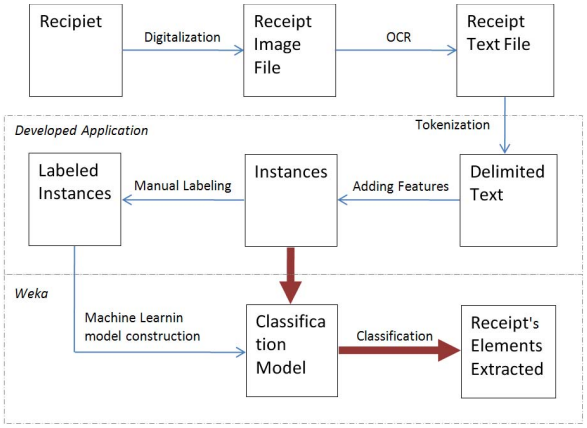

Figure 1. Approach for automated information extraction from receipts

2.1. Pre-processing

To automatically processes the receipt, firstly, it has to be converted into digital form; this is performed by scanning. Result of digitalization is typically an image. Next step is to convert the image of the receipt to a textual file, which can be performed using some OCR (optical character recognition) algorithm. Transforming the receipt into the digital text format is not in focus of the work, this process is well covered by the reference [2].

2.2. Generating datasets

Input is a textual file representing a receipt (type of paper receipt which is usually issued to private customers in the shops). Firstly, input textual file is delimited to words, based on empty spaces. Next, each word becomes a feature vector with 18 attributes that describe that word. To accomplish this, an application is developed. After textual file is loaded with the application, 17 values for the attributes are automatically assigned to each word (detailed description of the attributes is in chapter 3). The 18th attribute is

the target variable (class), a variable that is telling us what role has this word in the textual document. In our example the role of the word can be: company name, date of receipt, address of the shop, price of a product, product name, time of the purchase, unit for the product, unit price, or quantity. Initially, target variable cannot be automatically determined with the application. This will be accomplished with machine learning after the classification model is build (detail description in chapter 2.3). To build the classification model a set of labeled feature vector (i.e. words with already known class) is needed. The application enables labeling words by clicking the word of the displayed document and choosing one of the pre-defined possible classes for this word (figure 2 shows words in red squares which were labeled). After the set of labeled feature vectors is created in this way, classification model is built. With the classification model, roles of the words can be determined automatically with some accuracy (experimental results are given in chapter 4).

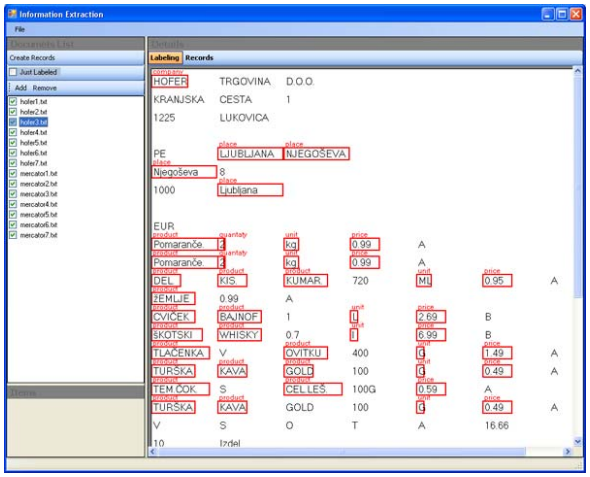

Figure 2. Labeling of a receipt using developed application

The main purpose of the application is generating feature vector sets from input textual documents. Currently, generated feature vectors can be outputted into textual or CVS file. This output enables experimenting with some machine learning applications (like Weka), to determine the performances of the approach and the best machine algorithm for this task. If the experiments show satisfying results after sufficient testing, this application can be integrated into the practical application for a real world problem. The example can be application for managing personal spending with scanning the receipts only. After few initial purchases from new shop, person would have to label the receipts manually, but later, elements of the receipt would be identified automatically. With this an overview of spending of a person (daily, weekly, etc.) can be given, with various possibilities, like: mostly used shop, average spending for different categories of products, the most expensive product, etc.

2.3. Using Machine Learning

As indicated with the figure 1, third part of proposed approach is dealing with creating classification model and classifying new unlabeled receipts. This part is conducted using Data Mining Open Source Machine Learning Software – Weka [3]. The supervised machine learning methods are used. This means that for building classification model, algorithm needs training samples which are already labeled. Weka is used to generate classification model, with labeled records generated using our software as an input. After the classification model is build, Weka software is used to classify unlabeled receipts based on previously build classification model. Several machine learning algorithms are tested, but best performing ones are decision tree and KStar (the performance are shown in chapter 4).

3. Feature set

Feature set is a set of variables that describe the elements of a receipt. Well defined feature set is crucial for performance of machine learning classification algorithms. Feature set for proposed system has 17 variables, plus one target variable (class). Variables of feature set are given in table 1.

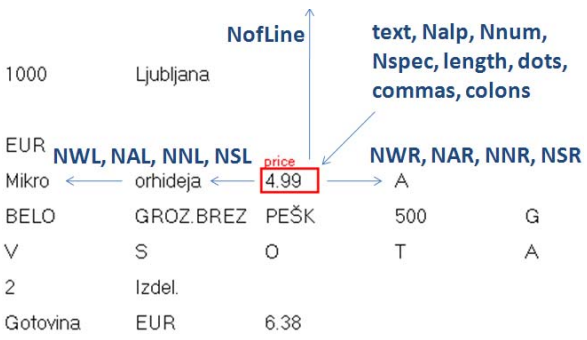

Figure 3. Variables of the feature set

The target variable is called class, it is of nominal type and it can have one of these values: *company*, *date*, *place*, *price*, *product*, *time*, *unit*, *unit_price*, *quantity*, or *other*.

Table 1. Elements of the feature set

variable	description
text	*text of the element*
Nalp	*number of alphabetic characters in the element*
Nnum	*number of numerical characters of the element*
Nspec	*number of special characters of the element*
length	*the length of the element*
dots	*number of dots in the element*
commas	*number of commas in the element*
colons	*number of colons in element*
NWL	*number to words left to the element*
NWR	*number of words right to the element*
NAL	*number of alphabetic characters left to the word*
NAR	*number of alphabetic characters right to the word*
NNL	*number of numeric characters left to the word*
NNR	*number of numeric characters right to the word*
NSL	*number of special characters left to the word*
NSR	*number of special characters right to the word*
NofLine	*line of the receipt (normalized to 1)*
CLASS	*role of the element on the receipt*

Figure 3 shows one part of the receipt with an element labeled as a *price*. As indicated with the figure, variables of feature set can be grouped into the three groups: (1) variables that are determining the content of the element (text, Nalp, Nnum, Nspec, length, dots, commas and colons); (2) variables that are determining the location of the element (NWL, NWR and NofLine); and (3) variables that are determining the content of elements' surrounding (NAL, NAR, NNL, NNR, NSL and NSR).

4. Experimental results

To test the performance of the proposed system, datasets were created using the designed application and classified using Weka.

Firstly, receipts were converted into digital textual format. Since digitalization and conversion of an image to a text is not a part of this research, conversion was done manually by retyping the receipts into a textual file. Fourteen receipts were digitalized. They were taken from two grocery stores of different companies. The receipts were from different buyers - with diverse lengths and products. Next, the receipts were loaded into developed application and labeled (as shown on figure 2). Finally, three different data sets were created: first (A) - with containing only receipts of purchases from shop *A* receipts, second (B) - with receipts of purchases from shop *B*, and third (AB) - with receipts of purchases from both companies *A* and *B*. The datasets A, B and AB have 600, 815 and 1415 instances respectively.

Weka was used to load the datasets and to perform different classification algorithms, in order to test the performance of the system. Ten folds cross validation technique was used for testing. Algorithms applied were: J48 decision tree and KStar instance based model.

J48 did not show the best results, but the classification model is highly interpretable. Figure 4 shows the part of the decision tree classification model used for dataset A.

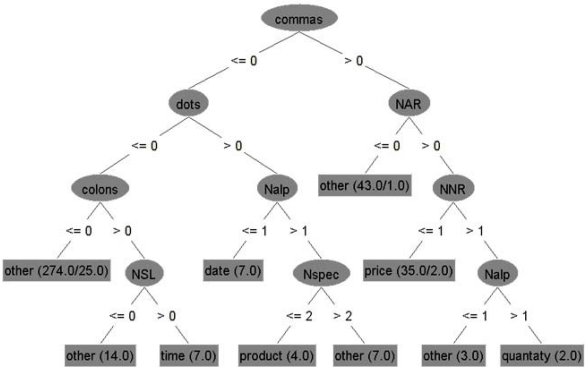

Figure 4. Part of the decision tree classification model

From this decision tree classification rules can be easily derived, for example, IF number of commas in the element is smaller or equal to zero AND number of dots is greater than zero AND number of alphabetical characters is

smaller or equal to 1 THEN the element is classified as a *date*.

KStar is an instance based classification model which uses entropy as a distance measure [4]. KStar model performed much better than other algorithms, e.g. Bayes classifier and SVM.

The classification accuracy after testing with J48 and KStar is shown in table 2.

Table 2. Classification accuracy of applied algorithms

dataset	J48	KStar
A	94.67%	97.83%
B	94.47%	94.36%
AB	92.16%	95.90%

The results in the table 2 show that in the practical application when user would try to automatically identify elements of the receipt of purchase from shop A, using classification model with only training receipts from shop A, this can be accomplished with accuracy 94.67% (using J48 algorithm). If the user would try to identify elements of purches from either shop *A* or *B*, using classification model build from trainig reciepts from both shops, this can be accomplished with accuracy 92.16% (using J48 algorithm).

5. Discussion

Approach for automated information extraction from receipts using Machine Learning was introduced. Developed application enables labeling of receipts and generating datasets. These are the basis for applying Machine Learning algorithms for classification. Experiments indicate potential usefulness of the approach, but the additional experiments are needed for more credible results. To make the testing more feasible, in the future work the proposed system will be integrated with digital character recognition system, so that the input textual files can be generated automatically with scanning, without need for manual retyping. The approach can be especially useful for semi-automatic systems, where user monitors the output and corrects the misclassification, training the system in the process. The possible practical application could be managing personal expenses. Approach was tested on the receipts, but it can be applied for any kind of textual documents.

6. References

[1] Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Commun. ACM* 39, 1 (January 1996), 80-91. DOI=10.1145/234173.234209 http://doi.acm.org/10.1145/234173.234209

[2] Chen Q., Evaluation of OCR Algorithms for Images with Different Spatial Resolutions and Noises; Master Thesis, University of Ottawa; 2003

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[4] John GC, Leonard ET: K*: An Instance-based learner using an entropic distance measure. the Proc. of the 12th International Conference on Machine learning 1995:108-114.