Document Categorization Based On OCR Technology: An Overview

DARKO ZELENIKA, JANEZ POVH, ANDREJ DOBROVOLJC
DataLab - Laboratory of Data Technologies
Faculty of Information Studies
Ulica Talcev 3, Novo Mesto
SLOVENIA
darko.zelenika@fis.unm.si, janez.povh@fis.unm.si, andrej.dobrovoljc@fis.unm.si

Abstract: - In this paper a brief overview of document categorization process is presented, with the focus on documents obtained by OCR (Optical Character Recognition) technology. Work of different authors from area of document categorization is described. Text obtained by OCR needs to be prepared in a way that categorization algorithms can use it to provide better categorization accuracy, thus such methods are introduced. A comparison of results of different categorization algorithms is shown. Most authors obtained the best results with SVM (Support Vector Machine) classifier. Two document categorization software programs are introduced, both commercial and open source. An invoice recognition project on which authors of this paper are working on is introduced.

Key-Words: - document categorization, document categorization algorithms, document categorization software, OCR, OCR errors, text representation methods, invoice recognition

1 Introduction

Document categorization refers to a process of grouping documents into predefined set of categories. The purpose of this paper is to bring an overview of document categorization process based on the OCR (Optical Character Recognition) technology. The OCR technology transforms scanned document images into editable textual documents. However, the result of OCR process is not perfect because it doesn't give the same result quality for all documents [8]. For high-quality scanned documents OCR will give almost perfect result, but sometimes a few errors may occur. When dealing with low-quality scanned documents OCR will produce inaccurate results [8]. Therefore, the question is: Do OCR errors affect document categorization accuracy based on document's content?

A number of papers explored this problem. In [16] authors showed that OCR errors have little or no effect on categorization when classifier based on Naive Bayes model is used. Agarwal et al. in [3] used Naive Bayes and SVM (Support Vector Machine) classifier for text categorization and came to conclusion that accuracy of text categorization won't change much as long as up to 40% of document's content is affected by noise. Vinciarelli in [27] used clean and nosy version of the same texts and compared the both. Results showed that loss in categorization performance of nosy text is insignificant even when average word error rate is around 50%. For document categorization Price and Zukas in [22] used Latent Semantic Indexing (LSI) algorithm and showed that this technique is highly resistant to noise in text that is generated by OCR. Guowei et al. in [12] showed that accuracy of text categorization was not reduced significantly until word recognition rate was reduced to 80%.

On the other side Hoch in [21] has found that nosy OCR text can affect text categorization especially when disturbed words are words which are significant for a specific document category. Also, Taghva et al. [15] in their research run automatic correction of noisy text over two OCR documents that couldn't be categorized and after automatic correction one document did get categorized correctly. Brooks at al. [5] showed that with their approach almost all OCR errors can be removed. They compared standard text block from noisy structured scanned document with small set of known corrected text. The noisy text block is replaced with the text which is most similar to it. Volk at al. [28] created an algorithm that compares the output of two OCR systems (Abbyy FineReader 7 and OmniPage 17) and showed that combination of two outputs leads to improved recognition accuracy and reduces OCR errors.

The main purpose of the paper is to present work of different authors from the area of document categorization, introduce and compare categorization algorithms, present document

categorization software programs and invoice recognition project.

An important task in text categorization is to prepare and represent text in a way usable by a classifier [24], and such methods are introduced in Section 2. Popular algorithms which are used for document categorization are shown in Section 3 while Section 4 brings comparison of results of introduced algorithms from many different authors. In Sections 5 and 6 two document categorization software solutions are introduced both commercial and open source. Section 7 introduces a document categorization project on which authors of this paper are currently working on. Section 8 concludes the paper.

2 Document representation methods

This section will introduce different methods which are used to prepare recognized OCR text for categorization. Idea behind these methods is to assist categorization algorithms to improve categorization accuracy. Among all methods only next three methods will be introduced:

Elimination of stop words. Stop words are eliminated to speed up categorization process. These are the most common English stop words: *the, is, at, which,* and *on*. Stop words are eliminated because they don't explain what a document is about. Each language has list of stop words which is created manually.

Lemmatization is the algorithmic process of determining the lemma for a given word [30]. In English language the word *work* can be used as *worked*, *working*, *works* etc., therefore the word *work* is lemma. Authors in [4] showed that when *lemmatization* is used for categorization of documents which are written in Basque categorization results were improved.

N-grams of characters. An *N-gram* is an N-character slice of a longer string [29]. For the string "*This is n-gram text*" with the N=4 the result of 4-gram will be: [This, his_, is_i, s_is, _is_, is_n, s_n-, _n-g, n-gr, -gra, gram, ram_, am_t, m_te, _tex, text]. In [29] authors introduced *n-grams* base approach which is ideal for document categorization of noisy text that is coming out of OCR process.

3 Document categorization algorithms

Document categorization is a problem in information technology when hundreds of electronics documents need to be categorized automatically. To do such task computers need to have special software tool with intelligent algorithms. Categorization can be also considered as supervised classification. To categorize documents we need a labeled set, i.e. a set of documents where we know the category that the document belongs to. These documents we split into the training and testing part. Training part is used to build the model which will be tested using testing documents. In unsupervised categorization (also called classification), assignment to groups is done without having any information about the group that the documents really belong to [13]. Usually performed document categorization algorithms are:

Vector Space Model – centroid algorithm. This is one of the simplest categorization algorithms. The tf-idf (term frequency - inverse document frequency) is a weighting factor often used in information retrieval and text mining [32]. The tf-idf assigns a weight to each word in a document. The weight will be the highest if a word occurs many times within a few documents. It will be lower if a word occurs fewer times or it occurs in many documents. It will be the lowest if a word occurs in almost all of training documents [7]. The tf-idf weighting scheme is often used in the vector space model together with cosine similarity to determine the similarity between two documents [13]. During the learning stage only the average feature vector for each category is calculated and set as centroid-vector for the category [6]. A new document is categorized by finding the most similar centroid-vector (class vector) to its feature vector (document vector). One of disadvantages of this algorithm is its performance when the number of categories is very large [6].

Linear support vector machine is a supervised learning algorithm used for categorization [6]. Based on set of training documents, each belonging to one of two categories, Linear SVM builds a model which will categorize new document into specific category. Built model is a representation of documents as points in space where documents of different categories are separated from each other by clear gap that is wide as possible. Therefore, the main idea of Linear SVM is to construct a hyperplane as the decision surface such that the margin of separation between two categories of documents is maximized [31]. Documents which are closest to the decision surface are called support vector. SVM is a very powerful algorithm and has outperformed others algorithms in several studies [6].

k-Nearest neighbor. In pattern recognition, the k -nearest neighbor algorithm (k-NN) is a method for categorizing objects based on closest training

examples in the feature space [13]. What makes this algorithm interesting is that skips the learning phase. The categorization is performed on set of the training documents. The new document is categorized based on the categories of its k nearest documents (neighbors). How close documents are to each other can be evaluated by measuring for instance the angle between two feature vectors or calculating Euclidian distance between the vectors [6]. Although Euclidian distance is probably the most commonly used distance function, other distance functions like Manhattan or Maximum can be also used [19]. An advantage of the k-nearest neighbor algorithm is its simplicity but it is really slow in process of categorization especially when the number of training documents is growing [13].

Naive Bayes classifier is the simplest instance of a probabilistic classifier. Therefore, categorization is based on probability which means that the output *Pr(C|d)* of this classifier represents the probability that a document *d* belongs to a category *C* [25]. To each word in a document probability is given based on number of its occurrence in that particular document [25]. A new document is categorized based on the highest probability. Naive Bayes performs very efficiency during categorization of new documents [6][25].

4 Results of categorization – comparison of algorithms

In previous section popular document categorization algorithms were proposed. This section will show comparison of proposed algorithms. Table 1 shows comparison results of categorization algorithms from nine authors [6]. Most of the authors used news stories from different datasets for comparison. Most commonly used data set is Reuters-21578 [17]. Authors in [24] used OCRed documents for categorization and obtained results are same as results of other authors. In most cases SVM outperforms the other algorithms. Only in one case Naive Bayes outperformed SVM and in some cases it performed almost equally well as SVM.

5 Document categorization with ABBYY - commercial

ABBYY FlexiCapture is data capture and document processing software. FlexiCapture as input can have many various scanned documents and based on OCR technology it has ability to recognize text and objects from documents. Using obtained results from OCR it can categorize documents into specific categories and export recognized data to a different file formats.

To identify document type FlexiCapture uses FlexiLayout which is an integral part of the FlexiCapture software. For each document type one FlexiLayout needs to be created. FlexiLayout describes: how to identify a document, what data need to be extracted and how to find this data. Therefore, FlaxiLayout represents a simple document classifier which has to have unique features of a specific document. Based on these unique features it has ability to recognize the right type of the document. One or many features can be used for document recognition. For example, to recognize an invoice the software can search for the keyword "*invoice*" or "*invoice number*". Except text i.e. keywords other features can be also used such as: bar codes, lines, logos and gaps. The categorization process with FlexiLayout is simple and takes very little time [2].

Authors	Centroid	SVM	k-NN	NB	Result
[1]	x	x	x		SVM performed better than other algorithms.
[9]	x	x		x	(SVMs) are very accurate, quick to train, and quick to evaluate.
[26]		x		x	SVM performs better then Naive Bayes.
[14]		x	x	x	SVMs consistently achieve good performance.
[33]		x	x	x	SVM and k-NN outperformed NB when there is less than 10 instances per category, but all algorithms performed similarly when there are more than 300 instances.
[23]		x	x		SVM performs better then K-NN.
[11]	x	x	x		SVM performs better than other algorithms.
[24]		x	x	x	Best results are obtained with the SVM.
[25]		x		x	Naive Bayes is the best in terms of accuracy and computational efficiency.

Table 1. – Comparison of different categorization algorithms from different authors [6]

ISBN: 978-960-474-304-9

FlexiCapture has three categorization modes: *autolearning, rules-based* and *combined*. Based on a set of training documents in the *autolearning* mode software automatically recognizes unique features and learns to recognize different document types.
The *rules-based* mode recognizes documents based on rules which are specified by the user. In *combined* mode both *autolearning* and *rules-based* modes are used: first *autolearning* mode is applied and in case of uncategorized documents *rules-based* mode is additionally applied [2].

Mikrografija d.o.o.[1] is Slovenian company which offers modern solutions for electronic document management and electronic archiving [18]. They are daily facing with a huge number of received documents of different types from their clients. One of solutions they offer is data capture solution "mZajem" which uses different OCR tools with ability to read and process the documents. One of OCR tools they already use is ABBYY FlexiCapture. They want to use this tool in order to automatically categorize documents with very high accuracy and export data to different systems (CRM, ERP, DMS,...). Since they want to provide such service for small and medium-sized enterprises from Slovenia and some other countries from Adriatic region, for a small price, they are highly motivated to compose new software based on available SDK.

6 Document categorization with Ephesoft – open source

Ephesoft is open source document capture software which can be trained to distinguish the difference between different types of documents and extract meta-data from the content of a document using OCR technology. Ephesoft has different categorization solutions that can be used depending on the format of scanned document image [20]. These solutions are:

Barcode categorization is categorization of documents based on special barcodes. To achieve this Ephesoft uses barcodes such as QR, 3 of 9, and Data Matrix [20].

Image/Layout categorization. For documents with unchanged structure and finite pages categorization can be achieved based on their layouts. Ephesoft uses this technique when categorization can't be done based on documents content. Examples of these documents are fixed forms which do not have a lot of text or their text content is unpredictable but their physical appearance i.e. layout, graphics and formatting, is consistent [20].

Extraction based categorization. By using OCR output documents can be categorized based on the keywords. Ephesoft can be configured to capture unique feature from a document like invoice numbers so that documents like this can be categorized as invoices [20].

Content analysis based categorization. Documents can also be categorized by using a few documents as training sample. Based on training sample Ephesoft learns the content of the document and creates model for that type of document. Each model represents a document category. A new document will be a part of a specific category only if its content is similar to a model of that category [20].

As an example we mention Mountain West Financial, Inc. which is a full-service, privately held mortgage banker that processes a tremendous volume of paperwork. All this paperwork was processed manually before using Ephesoft solution. With Ephesoft they trained 225 mortgage document types. Mountain West is reporting 95% accuracy in document categorization process [10].

7 Invoice recognition project

The general activity of Mikrografija d.o.o.[1] is to capture, process and storage paper documents into electronic form. The problem the company faces is that small and medium-sized enterprises have hard time in making decision when it comes to using automatic data capture software which will convert their paper documents into electronic ones. The main reason of this problem is the price of such software. On the other side, what most of enterprises want to only have is automatic capture of received invoices. Therefore, they don't need complex data capture software tools like TIS, Kofax and ABBYY.

Mikrografija d.o.o.[1] wants to offer an invoice recognition software solution to the enterprises from Adriatic region with the reasonable price which will be able to train, recognize and extract data from invoices. Such software will be able to recognize special characters like (č,ž,š) which are present in Adriatic region. Faculty of Information Studies[2] is doing this project for Mikrografija d.o.o.[1] Researchers from Faculty of Information Studies need to explore the market to find needs of enterprises from Adriatic region and develop such software by using ABBYY SDK. We will report the results on the conferences in 2014.

1 Mikrografija d.o.o. - http://www.mikrografija.si/
2 Faculty of Information Studies - http://www.fis.unm.si/si/

ISBN: 978-960-474-304-9

8 Conclusion

In this paper basic steps of automatic document categorization based on OCR technology are introduced. The quality of scanned document is not always good enough, so document categorization is usually done over noisy documents' content. Work of different authors is introduced. Some authors showed that a little noise can't affect document categorization accuracy much. On the other side some of them showed that noise can affect document categorization accuracy if disturbed words are significant for specific category. Some authors introduced different solutions for noise reduction. An important step in document categorization is text representation and preparation. Result of text representation and preparation methods is handed over to document categorization algorithm. Couple of algorithms are introduced and compared. SVM usually performs the best, while Naive Bayes often performs almost well as SVM. Document categorization solutions of commercial ABBYY FlexiCapture and open source Ephesoft software programs are explained.

Automatic document categorization software is very helpful and time-saving for enterprises. Small and medium-sized enterprises are mostly dealing with invoices and often refuse such software because of its complexity and price. Authors of this paper are part of the project whose aim is to develop affordable invoice recognition software for small and medium-sized enterprises in Adriatic region.

Acknowledgment:
Work supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence.

References:

[1] A. Cardoso-Cachopo and A. L. Oliveira, An empirical comparison of text categorization methods, *In Proceedings of SPIRE-03, 10th International Symposium on String Processing and Information Retrieval*, pages 183–196, Springer Verlag, 2003.

[2] ABBYY, Insights into Abbyy FlexiCapture, Available from: http://www.abbyy.de/flexicapture-technologie-broschuere, (27.04.2013).

[3] Agarwal S., Godbole S., Punjani D., Shourya Roy, How Much Noise Is Too Much: A Study in Automatic Text Classification, *Seventh IEEE International Conference on Data Mining*, 2007, ICDM 2007.

[4] Ana Zelaia, Iñaki Alegria, Olatz Arregi, Ana Arruarte Lasa, Arantza Díaz de Ilarraza Sánchez, Jon A. Elorriaga, Basilio Sierra, Exploring Basque Document Categorization for Educational Purposes using LSI, *CSEDU (1)* 2009, p.p. 5-10

[5] Brooks R., Hunnisett D. and Teahan W.J., "A practical implementation of automatic text categorization and correction for the conversion of noisy OCR documents into Braille and large print", *Workshop on Noisy Data, IJCAI'2007, Workshop on Analytics for Noisy Unstructured Text Data*, 2007.

[6] Brucher H., Knolmayer G., Mittermayer M.A., Document Classification Methods for Organizing Explicit Knowledge. In: *Proceedings of the Third European Conference on Organizational Knowledge, Learning, and Capabilities*, Athens, Greece (2002)

[7] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, *Cambridge University Press*, 2008, p.p. 119.

[8] CVISION technologies inc., OCR for Poor Quality Scan, Available from : http://www.cvisiontech.com/reference/ocr/ocr-for-poor-quality-scan.html?lang=eng (29.04.2013)

[9] Dumais S., Platt J., D.H., Sahami, M., Inductive learning algorithms and representations for text categorization. In: *CIKM*, ACM (1998), p.p. 148–155.

[10] Ephesoft, Mountain West Financial leverages Ephesoft's Advanced Imaging Software to streamline mortgage loan processing. Available from: http://www.ephesoft.com/images/stories/ephesoft/MtWest_Financial_Case_Study_August_2011.pdf (27.04.2013)

[11] Goller C., Löning J., Will T., Wolff W., Automatic Document Classification: A thorough Evaluation of Various Methods, *In Knorz, G., Kuhlen, R. (eds.): Informationskompetenz-Basiskompetenz in der Informationsgesellschaft. Proceedings 7. Intl Symposium für Informationswissenschaft*, (ISI 2000), pp. 145–162.

[12] Guowei Zu, Mayo Murata, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura, The Impact of OCR Accuracy on Automatic Text Classification, *AWCC* 2004, pp. 403-409.

[13] Institute for Information Systems and Computer Media., E-Learning Course: Document Classification, Available from: http://www.iicm.tugraz.at/cguetl/courses/isr/opt/classification (24.4.2013)

[14] Joachims, T., Text Categorization with Support Vector Machines: Learning with Many Relevant Features, In: *Proceedings of the 10th European Conference on Machine Learning*, 1998, pp. 137-142.

[15] Kazem Taghva, Thomas A. Nartker, and Julie Borsack, Recognize, Categorize, and Retrieve. *In Proc. of the Symposium on Document Image Understanding Technology*, pages 227–232, Columbia, MD, April 2001, Laboratory for Language and Media Processing, University of Maryland.

[16] Kazem Taghva, Thomas A. Nartker, Julie Borsack, Steven Lumos, Allen Condit, and Ron Young, Evaluating text categorization in the presence of ocr errors. *SPIE*, 2000.

[17] Lewis, David (1997) Reuters-21578 Reuters-21578 Text Categorization Test Collection Distribution 1.0. Available from: http://www.daviddlewis.com/resources/testcollections/reuters21578/

[18] Mikrografija d.o.o, "Zajem podatkov - mZajem". Available from: http://www.mikrografija.si/ (29.04.2013)

[19] Parvin H., Alizadeh H., Minaei-Bidgoli B., A New Divide and Conquer Based Classification for OCR, *In M. Crisan (Ed): Convergence and Hybrid Information Technologies; I-Tech Education and Publishing*, 2010; ISBN: 978-953-307-068-1.

[20] Pat Myers, Ike Kavas, Michael Muller and Clifford Laurin, *Intelligent Document Capture with Ephesoft*, Packt Publishing, 2012

[21] Rainer Hoch, Using IR techniques for text classification in document analysis, *In SIGIR '03*, pages 104–110, ACM Press, 1994.

[22] Robert J. Price and Anthony Zukas, Accurate document categorization of OCR-generated text. *In Symposium on Document Image Understanding Technology*, 2005.

[23] Siolas, G., d'Alche-Buc, F., Support Vector Machines based on a semantic kernel for text categorization, *In: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00),* pp. 205-209, 2000.

[24] S. Laroum, N. Béchet, H. Hamza, M. Roche, Hybred: An OCR document representation for classification tasks, *In International Journal of Data Engineering and Management (IJDEM)*, Vol 1, 2010.

[25] S.L. Ting, W.H. Ip, Albert H.C. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification?", *In: International Journal of Software Engineering and Its Applications*, Vol. 5, No. 3, July, 2011, pp. 37-46.

[26] Tong Zhang , Frank J. Oles, Text Categorization Based on Regularized Linear Classification Methods, *Information Retrieval*, V.4 N.1, p.5-31, April 2001.

[27] Vinciarelli, A., "Noisy text categorization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*,Volume 27 (12), 2005.

[28] Volk M., Marek T., Sennrich R., Reducing OCR errors by combining two OCR systems. In: *ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, Lisbon, Portugal, 16 August 2010 - 16 August 2010, pp. 61-65.

[29] W.B. Cavnar and J.M. Trenkle, N-Gram-Based Text Categorization, *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161--175, Las Vegas, US. (1994).

[30] Wikipedia: The Free Encyclopedia (2013), "Lemmatisation". Available from: http://en.wikipedia.org/wiki/Lemmatisation (29.04.2013)

[31] Wikipedia: The Free Encyclopedia (2013), "Support Vector Machine". Available from: http://en.wikipedia.org/wiki/Support_vector_machine (29.04.2013)

[32] Wikipedia: The Free Encyclopedia (2013), "tf-idf" . Available from: http://en.wikipedia.org/wiki/Tf%E2%80%93idf (29.04.2013)

[33] Yang Y., Liu X., A Re-Examination of Text Categorization Methods, *in: Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 42-49, 1999.