ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Model for Document classification using Kernel Discriminant Analysis(KDA) and semantic analysis

Chirag Patel
Smt. Chandaben Mohanbhai Patel Institute of Computer Applications
CHARUSAT
Changa, Gujarat, India

Mahesh Gadhavi
Smt. Chandaben Mohanbhai Patel Institute of Computer Applications,
CHARUSAT
Changa, Gujarat, India

Abstract: In recent days, digital communication has become inevitable. The scope of this research is to provide a theoretical framework of automatically classification of question paper. It can be helpful to the library of any University to automate their archival process. A survey has been conducted to understand the existing systems in this research area. After doing the survey we observed that many authors have done significant work in document classification but little work has been done to automatically store the documents in particular folder. Therefore, there is a huge scope to develop working model of the framework suggested in this paper.

Keywords: OCR, Kernel Discriminant Analysis (KDA), Semantic Analysis, document classification, binarization

I. INTRODUCTION

In the era of digital word, storing printed documents in digital form has become necessary evil. Many printed or hand written documents are available in library but due to lack of maintenance these documents are not preserved in the computerized form. Due to this problem the documents are not available in the digital form to the target audience. The classification and storing the digital copy of question papers can be challenging task for a librarian in the academic institutions. For such situations it is necessary to develop a system which can automatically scan the question paper, classifies the branch, subject and semester details and store it into appropriate folder. In this paper we have proposed automatic document classification method to help librarian in managing the digital copy of question paper. The Kernel Discriminant Analysis (KDA) is used to do OCR for converting printed text into ASCII text. The KDA is pre-processing step to do document classification. For document classification semantic analysis is applied as it is fast efficient. In the following section, a literature review is conducted to study and analyse present state of work done in this area. The model is well explained in section III. The paper is concluded in the conclusion section.

II. RELATED WORK

Managing multiple documents and classification of such documents is the cumbersome task. Some of the authors made an attempt of classify the semi structured XML document [7]. Classification can also be used to understand the similarity [10] between two documents. An optimization scheme to enhance class discriminant is proposed by [6]. They focused mainly on non-linear classification. Their algorithm optimizes the algorithm and by reducing dimensionality of feature space. The authors claim to have better performance as compared to standard problem. For multi-label document classification ensemble based approach is proposed by [16]. They applied Naive Bayes and logistic regression methods for classification. The authors mention that in multi-label classification, topology is not an ensemble for Binary Relevance (BR) classifier. They applied four methods mainly: batch, incremental, ensemble, and ensemble incremental to obtain better results in their experiments. Another multi-label document classification based on statistical topic model is proposed by [15]. They applied three different approaches like flat-Linear Discriminant Analysis (LDA), prior LDA and dependency LDA to achieve good classification results. Another LDA based approach is mentioned in [5]. Today, we have huge amount of unstructured data available around us. To classify documents from unstructured data [12] proposed a policy-driven framework. The authors mention that Natural Language Processing (NLP) helps a lot for text processing. They applied four steps based method. The first step is sensitivity level classification which divides the resources in three categories: high, medium and low. The second step, SystemT information extraction is used to extract information from the resource document using proprietary tool of IBM SystemT [17]. This tool is used to extract structure data from unstructured or semi-structured document. Annotation Query Language(AQL) as language for data extraction in this tool. In final frequency analysis i.e. the third step, the frequency of keywords is calculated by using SystemT. In the final step i.e. policy engine, the sensitivity level of the documents is prepared by comparing the values defined in the already defined policy. In recent year we receive many spam e-mail in out mail box. To decide whether to e-mail is spam or not, hybrid decision tree and logistic regression based spam classifier is proposed by [21]. The applied hybrid method called LRFNT+DT which uses combination of Logistic Regression (LR) and Decision Tree (DT) to identify spam mail. The LR method is used to remove noisy data. Then the cleaned data is forwarded to DT for classification. They achieved 91.67% of accuracy using this hybrid method. For document classification effective method is developed by [19]. The authors proposed new method known as Labeled Dirichlet process mixture models of von Misesâ˘A ¸ SFisher distributions (LDPV) for text classification. After applying training this model, they

tested this model to achieve comparable level of accuracy. In our research major task is to automatically copy the question paper in the related folder. Similar work has been proposed by [8]. The method proposed by the authors is based on Bayesian Support Vector Machine (SVM). They applied SVM classifier at back end and Bayesian classifier at the front end to obtain promising results on their experiments. They achieved 80.33% of overall accuracy by using these hybrid methods. Another SVM based approach is proposed by [9]. The authors applied Euclidean distance based approach for automatic text document categorization. The experiments were carried out Intel Core i3 process having 3.2 speed of GHz 2 GB RAM and Windows 7 operating system to have 84.73% of overall accuracy. For real-time applications, a business document classification approach is proposed by [3]. The graph coloring is used for layout analysis and document classification. A K-nearest neighbour (KNN) based approach is discussed in [2]. The authors applied this method for document classification. The authors mention that KNN rule based methods are widely used methods for document classification. In digital devices such as scanner or camera, automatic document classification can be handy. To do such task low-entropy pipelines base approach is proposed by [11]. They applied halftone classifier to achieve 97% for accuracy in the environment of Intel Core i7 with 3.40 GHz computer. In [14], different approaches of document classification for electronic document management system are discussed. Apart from the literature discussed here there are other good approaches for document classification such as multi instance stream learning framework [1], word spotting technique from document image [4], Ensemble of keyword extraction methods [13], clustering and novel HMM based method [20].

It is evident that there are numerous methods available for document classification but little work has been done to automatically classify and store document in particular folder. So there is a huge scope to develop such system. In the following section theoretical framework of such system is proposed.

III. PROPOSED WORK

In this research we propose a solution to automatically classify and store the scanned question paper by using KDA and semantic analysis. The architecture of the system is presented in Figure 1. First the scanned copy of question paper is provided as input to system. Then it must be converted to binary image and the binarized document to the OCR engine. The image binarization process used to convert any image to black and white image. Several methods are available for image binarization [18], [22]. The OCR engine converts the printed text of question paper into ASCII by

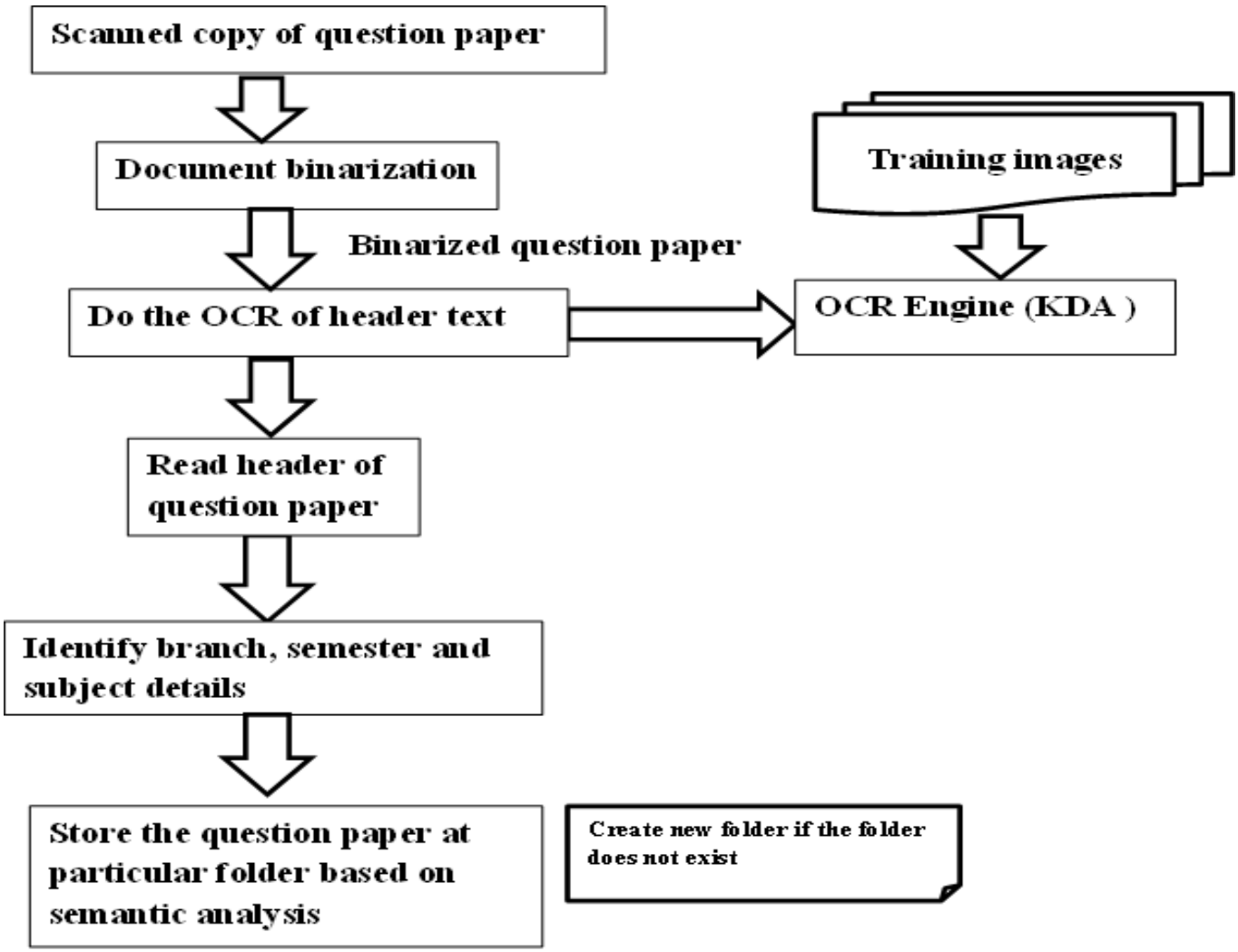

Fig.1. System Architecture

using the KDA. The KDA is trained using large number of training images in the ASCII form. The first few lines of header of the question paper are needed to classify and store the question paper in particular folder. This is done by the semantic analysis rules which help to store paper in related subject folder. The KDA and semantic analysis methods are well explained in following sub sections.

A. KERNEL DISCRIMINANT ANALYSIS (KDA)

The kernel discriminant analysis (KDA) is the method based on the statics and conditional probability to do nonlinear multi-class classification. It is modified version of Linear Discriminant Analysis (LDA)- which is used to do binary classification. The method is explained in equation 1.

$$\alpha_i = \frac{1}{e_i} \sum_{n=1}^{e_i} K_n^i \quad \ldots\ldots\ldots\ldots \quad (1)$$

Here e_i contains the number of examples belong to class m. The classification process is based on the maximal likely-hood of the mean and co-variance. To apply KDA for training, numbers of training images are fed into the model.

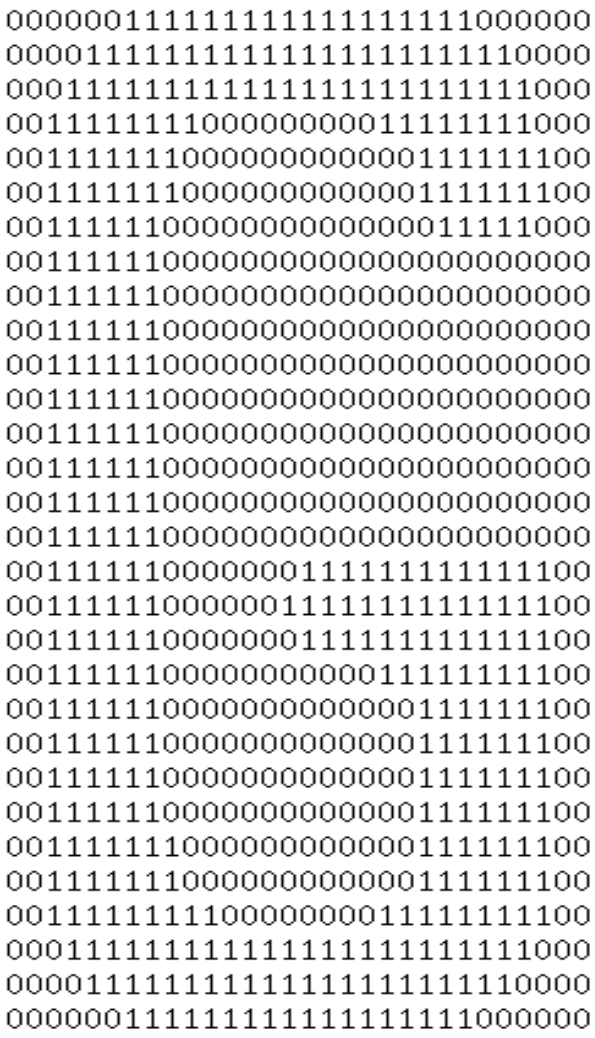

Fig.2. Training image example

These images are used to train the KDA to calculate maximal likely-hood of the predicted class. The sample of training image of digit 6 is shown in Figure 2.

To classify and store the question paper in related folder, semantic analysis method is employed. The semantic analysis is very popular method in Natural Language Processing (NLP) domain. It is well explained in following section.

B. SEMANTIC ANALYSIS

The semantic analysis is a part of Natural Language Processing (NLP). It includes understanding of the meaning of word, phrase, statement or paragraph. The semantic analysis should extract various words from the header of the question paper, extract the meaning and store the question paper in the related folder. The steps are mentioned below:

Step 1: Read the header line of the question paper.
Step 2: Extract University, Institute, Subject, semester and year from the header.
Step 3: Store the question paper in University\ Institute\Semester\Subject\Year
Step 4: If the folder is not created then create it and go to step 3.

IV. CONCLUSION

In this paper, a theoretical framework for automatic document classification of scanned question paper is proposed. We have done literature review of existing systems available for document classification. It is observed that the framework proposed in this paper will provide better accuracy and will take less processing time.

V. ACKNOWLEDGEMENTS

The authors would like to thank Mr. Suresh Solanki librarian of CMPICA for providing number of question papers for study and analysis of question papers.

VI. REFERENCES

[1] A multiple-instance stream learning framework for adaptive document categorization. Knowledge-Based Systems, 120:198 – 210, 2017.
[2] Tanmay Basu and C. A. Murthy. Towards enriching the quality of k-nearest neighbor rule for document classification. International Journal of Machine Learning and Cybernetics, 5(6):897–905, 2014.
[3] Djamel Gaceb, Véronique Eglin, and Frank Lebourgeois. Classification of business documents for real-time application. Journal of Real-Time Image Processing, 9(2):329–345, 2014.
[4] Angelos P. Giotis, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. A survey of document image word spotting techniques. Pattern Recognition, 2017.
[5] Li He, Yan Jia, Zhaoyun Ding, and Weihong Han. Hierarchical classification with a topic taxonomy via lda. International Journal of Machine Learning and Cybernetics, 5(4):491–497, 2014.
[6] A. Iosifidis, A. Tefas, and I. Pitas. Enhancing class discrimination in kernel discriminant analysis. In 2015 IEEE International Conference on Acoustics, Speech and

Signal Processing (ICASSP), pages 1926– 1930, April 2015.

[7] M. Khabbaz, K. Kianmehr, and R. Alhajj. Employing structural and textual feature extraction for semi structured document classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6):1566–1578, 2012.

[8] Lam Hong Lee, Rajprasad Rajkumar, and Dino Isa. Automatic folder allocation system using bayesian-support vector machines hybrid classification approach. Applied Intelligence, 36(2):295–307, 2012.

[9] Lam Hong Lee, Chin Heng Wan, Rajprasad Rajkumar, and Dino Isa. An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization. Applied Intelligence, 37(1):80–99, 2012.

[10] Y. S. Lin, J. Y. Jiang, and S. J. Lee. A similarity measure for text classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 26(7):1575–1590, 2014.

[11] N. D. Lyfenko. Automatic classification of documents in a natural language: A conceptual model. Automatic Documentation and Mathematical Linguistics, 48(3):158–166, 2014.

[12] E. Nwafor, P. Chowdhary, and A. Chandra. A policy-driven framework for document classification and enterprise security. In 2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pages 949–953, 2016.

[13] AytuÄ§ Onan, Serdar KorukoÄ§lu, and Hasan Bulut. Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, 57:232 – 247, 2016.

[14] I. E. Paramonova. Electronic document-management systems: A classification and new opportunities for a scientific technical library. Scientific and Technical Information Processing, 43(3):136–143, 2016.

[15] Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. Machine Learning, 88(1):157–208, 2012.

[16] D. G. Saputra and M. L. Khodray. An ensemble approach to handle out of vocabulary in multilabel document classification. In 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), pages 1–6, 2016.

[17] IBM Web Site.

[18] Morteza Valizadeh and Ehsanollah Kabir. Binarization of degraded document image based on feature space partitioning and classification. International Journal on Document Analysis and Recognition (IJDAR), 15(1):57–69, 2010.

[19] Ngo VanÂ˜aLinh, Nguyen Kim Anh, Khoat Than, and Chien Nguyen Dang. An effective and interpretable method for document classification. Knowledge and Information Systems, 50(3):763–793, 2017.

[20] A. Seara Vieira, L. Borrajo, and E.L. Iglesias. Improving the text classification using clustering and a novel hmm to reduce the dimensionality. Computer Methods and Programs in Biomedicine, 136:119 – 130, 2016.

[21] A. Wijaya and A. Bisri. Hybrid decision tree and logistic regression classifier for email spam detection. In 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), pages 1–4, Oct 2016.

[22] Hosub Yoon Jaeyeon Lee Youngwoo Yoon, Kyu-Dae Ban and Jaehong Kim. "Best combination of binarization methods for license plate character segmentation". "ETRI Journal".