SCANDROID: AUTOMATIC CLASSIFICATION OF DOCUMENT IMAGES ON ANDROID MOBILE DEVICES

Gunjan Joshi*

Prateek Bedmutha*

Aman Patel*

Kinjal Bathani*

Abstract

Text categorization refers to the automatic labelling of documents, based on natural language text contained in or associated with each document, into one or more pre-defined categories. Today, image categorization is a necessity due to a very large amount of image documents that we have to deal with daily. The current image categorization system uses an associated text approach for classification of images. We propose herein a new approach for automatic image categorization on android mobile devices, an application for classification of document images based on its contents, which is useful to businessmen, teachers and students. The classification module is the primary module. OCR technology is used to extract the textual contents from the input images. The textual contents extracted are given as input to the classification module which automatically classifies the images based on hashing techniques. The searching module is used to search for relevant image documents based on user keyword. The interface of the Android OS makes the end-user easy and efficient to search the relevant images into the database based on user keyword. Our literature survey leads to conclusion that mining is a good and promising strategy for automatic image categorization.

Keywords— *Text Mining, Information Retrieval, Data Mining, OCR, Text classification, Feature extraction*

* Department of Computer Engineering, Sinhgad College of Engineering, University of Pune, India

I. INTRODUCTION

Our objective is to utilize the visual capabilities of the Android mobile phone to extract text from an input image. We use the camera features of the Android to capture image.

Any camera image of the document would be subject to several environmental conditions, such as variable lighting, reflection, rotation, and scaling (we would desire the same data to be extracted from the images regardless of the distance from the camera), among others.

Text classification attempts to associate a text with a given category based on its content. Text categorization is the task of automatically sorting documents into categories from a predefined set. In our work, we propose an evaluation of more significant amount of features.

Text Mining (TM) is a new, challenging and multi-disciplinary area, which includes spheres of knowledge like Computing, Statistics, Predictive, Linguistics and Cognitive Science. TM has been applied in a variety of concerns and applications. Some applications are summary creation, clustering, language identification, term extraction and categorization, electronic mail management, document management, and market research with an investigation.

TM consists of extracting regularities, patterns, and categorizing text in large volume of texts written in a natural language; therefore, NLP is used to process such text by segmenting it into its specific and constituent parts for further processing.

II. PROBLEM FORMULATION

A. *Problem Definition –*

Given an input textual image to the system, the system uses a web service to extract the textual features from the image which are used to auto classify the images. Data mining techniques are used for the same. The classified images are now easy to search for the given input keyword.

The relative images and the respective text is retrieved for the end-user as an end-result.

B. *Objective –*

The main objective of ScanDroid system is to provide the user with an efficient image retrieval tool that will allow the end-user to search through the database which is based on the content of the input textual image.

III. MATHEMATICAL MODEL

We now provide a model of the system in terms of Set theory domain.

1. Let 'S' be the Content based classification and retrieval of textual images.

   S= {……

2. Identify the inputs as I.

   S = {I…

   Where I = {I1, I2}

   I1 = {i | 'i' is the image file format from which text can be extracted} = {*.jpg,*.bmp,*.tiff,*.png}

   I2 = {t | 't' is location of input image file on the phone memory}

3. Identify the processes as P.

   S= {I, O, P …

   P= {Ex, Fe, Cf, R}

4. Ex is the set for extraction module activities.

   Ex = {Ei, Ep, Eo}

- Ei= {f| 'f' is the valid image file to extraction module.}
- Ep= {f | 'f' is the extraction function to convert the Fi to Fo.}

  Ep (Ei) = Eo
- Eo= {f | 'f' is the output generated by extraction module i.e. text document file}

5. Fe is the set for feature extraction module activities.

   Fe = {Fi, Fp, Fo}

- Fi= {f| 'f' is the valid input text document to feature extraction module.}
- Fp= {f | 'f' is the feature extraction function to convert the Fi to Fo.}

  Fp (Fi) = Fo
- Fo= {f | 'f' is the output generated by feature extraction module i.e. character sequence.}

6. Cf is the set of classification module activities.

   Cf = {Ci,Cp,Co}

- Ci= {f| 'f' is the features extracted to classification module.}
- Cp= {f | 'f' is the classification function to convert the Ci to Co.}

  Cp (Ci) = Co

- Co= {f | 'f' is the output generated by classification module.}

7. R is the set of search and retrieval activities and associated data.

   R= {Ri, Rp, Ro}

- Rip= {r| 'r' is the keyword query}
- Rp= {r| 'r' is searching and retrieval function}

  Rp (Ri) = Ro

- Ro= {r| 'r' is metadata of matching image file.}

8. Identify failure cases as F

   S= {I, O, P, F…

   Where F = {F1, F2, F3}

   Failure occurs when –

- F1= {l| 'l' is the image containing no text}
- F2= {p| 'p' is no matching keyword in domain dictionary}
- F3= {m|'m' is improper image}

9. Identify success case (terminating case) as E

   S= {I, O, P, F, E…

   Where E = {E1, E2, E3}

   Success is defined as-

   E1= {p| 'p' is image retrieved properly based on keyword.}

   E2= {q| 'q' is properly classified}

   E3= {r| 'r' is text file created successfully}

Mathematical Representation

Let 'S' be the system –

S= {I, Ei, Fi, Ci, Ri, Ep, Fp, Cp, Rp, Eo, Fo, Co, Ro, F, E}

where,

I = set of valid input data set where I = {I1, I2}.

I1 = set of valid image formats.

I2 = set of image locations on phone memory.

Ei = set of valid image files to text extraction module.

Fi = set of valid input text document to feature extraction module.

Ci = set of features extracted to classification module.

Ri = set of input to image retrieval module.

Ep = set of text extraction module function.

Fp = set of feature extraction module function.

Cp = set of classification function to convert the Ci to Co.

Rp = set of image retrieval module functions.

Eo = set of output of text extraction module.

Fo = set of output of feature extraction module.

Co = set of output generated by classification module.

Ro = set of output of image retrieval module.

F = set of failure cases.

E = set of success case.

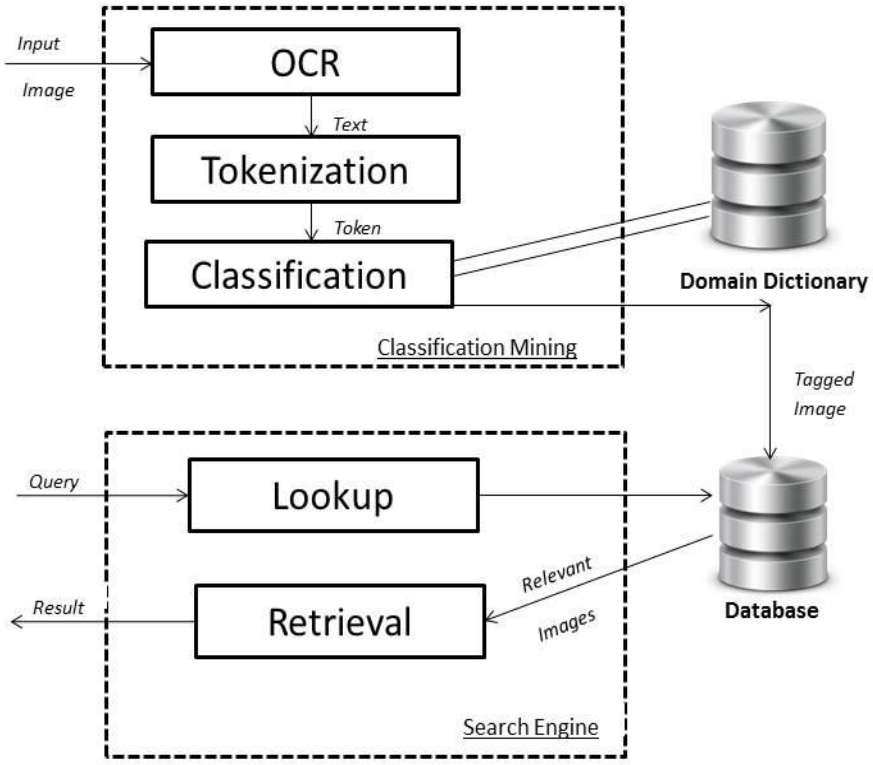

Fig 1. : Basic Flow Diagram of ScanDroid App

IV. FRAMEWORK

The basic framework for ScanDroid system consist of major two categories namely Classification and Retrieval.

1. Classification

It includes 3 phases, the end-result is classified image.

- Phase 1 (Pre-processing Phase) is the text extraction processing. OCR is used for the same.

- Phase 2 (Tokenization Phase) includes the creation of tokens which is required for classification. This process includes the removal of stop words and filtering of the words which is known as stemming.

- Phase 3 (Classification Phase) is responsible for assigning categories based on the keywords from feature extraction module.

2. Retrieval

It includes 2 phases, the end-result is the respective image.

- Phase 1 (Lookup Phase) takes an input keyword, compares with all the indexed items in the table to get the relevant matches.

- Phase 2 (Retrieval Phase) analysis all the matched keywords and returns the respective image(s).

V. ALGORITHM

There are two algorithms followed in the application namely: The Naïve Bayes classifier algorithm and the Searching Algorithm

- Naïve Bayes Algorithm

The naive Bayes classifier has been successfully used in the Rainbow text classification system. Let C = (c1,….,c_m) be m document classes. Given a new unlabelled document D and its corresponding word-list $\bar{W}$ = (ω 1, …., $\omega_{d'}$) (defined in the same way as the word-list for the training set), the naïve Bayes approach assigns D to a class c^*_{NB} as follows:

$$c^*_{\mathrm{NB}} = \mathrm{argmax}_{c_j \in C} P(c_j) \prod_{i=1}^{d'} P(w_i|c_j),$$

Where $P(c_j)$ *is the priori* probability of class c_j and $P(\omega_i \mid c_j)$ is the conditional probability of word ω_i given class c_j, the probabilities of words occurring in a document are independent of each other.

When the size of the training set is small, the relative frequency estimates of probabilities, $P(\omega_i \mid c_j)$ will not be reasonable; if a word never appears in the given training data, its relative frequency estimate will be zero.

The estimate of the probability $P(\omega_i \mid c_j)$ is given as:

$$P(w_i|c_j) = \frac{n_{ij} + 1}{n_j + k_j}$$

where n_j is the total number of words in class c_j, n_{ij} is the number of occurrences of word ω_i in class c_j and k_j is the vocabulary size of class c_j. This is the result of the Bayesian estimation with a uniform prior assumption, i.e. probabilities of the occurrence of words appearing in class c_j are equally likely.

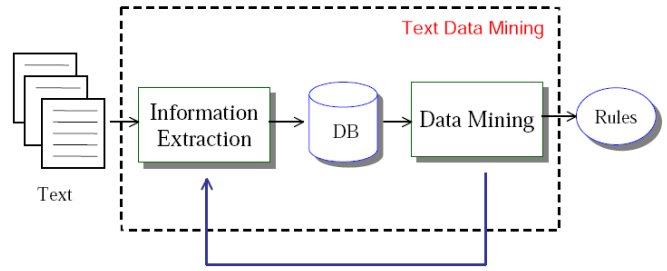

Fig 2: Overview of Text Mining

- Content based searching algorithm

The searching algorithm uses the searching within the hash table. Search query is given input to the system. The query is compared with the key value of the hash table. The hash table contains all the extracted documents and the relevant text. The files containing the query are returned to the list view of the GUI. The user has the ability to view the retrieved files.

VI. EXPECTED RESULT

The expected output of the system is a set of classified images retrieved after text mining techniques are applied on the textual images.

Another expected output is the respective image(s) based on the keyword given as input to the system.

The user may click the text file which is retrieved along with the image to view the textual contents of that particular image file.

VII. CONCLUSION

Therefore, we have presented a framework to perform Automatic Classification of Document Images on Android Mobile Devices. We have also formulated and described the mathematical model for the same.

ACKNOWLEDGMENT

This research paper cannot be considered complete without mentioning **Prof. N.G.Bhojne**. We wish to express true sense of gratitude towards his valuable contribution. We are grateful to him for his constant encouragement and guidance in the fulfilment of the project activity.

REFERENCES

[1] Dr.R.Geetha Ramani, G. Sivagami, Shomona Gracia Jacob. *"Feature Relevance Analysis and Classification of Parkinson Disease Tele-Monitoring Data Through Data Mining Techniques"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3, March 2012, ISSN: 2277 128X.

[2] Shalini Puri. *"A Fuzzy Similarity Based Concept Mining Model for Text Classification"*, IJACSA, Vol. 2, No. 11, 2011.

[3] Vishal Gupta, Gurpreet S. Lehal. *"A Survey of Text Mining Techniques and Applications"*, Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, August 2009.

[4] Atika Mustafa, Ali Akbar, and Ahmer Sultan. *"Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization"*, International Journal of Multimedia and Ubiquitous Engineering,Vol. 4, No. 2, April, 2009.

[5] M.Ikonomakis, S.Kotsiantis, V.Tampakas. *"Text Classification Using Machine Learning Techniques", WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp. 966-974.*

[6] S. M. Kamruzzaman, Farhana Haider, Ahmed Ryadh Hasan. *"Text Classification Using Data Mining"*, ICTM 2005.

[7] Raymond J. Mooney, Un Yong Nahm. *"Text Mining with Information Extraction", Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.*

[8] Dan S. Bloomberg, Gary E. Kopec and Lakshmi Dasari, *"Measuring document image skew and orientation"*, IS&T/SPIE EI'95, Conference 2422: Document Recognition II, pp. 302-316, Feb 6-7, 1995, San Jose, CA.

[9] Sami Laroum, Nicolas Béchet, Hatem Hamza, and Mathieu Roche. *"HYBRED: An OCR Document Representation for Classification Tasks".*

[10] Sonia Bhaskar, Nicholas Lavassar, Scott Green. *"Implementing Optical Character Recognition on the Android Operating System for Business Cards"*, EE 368 Digital Image Processing.

[11] Ray Smith, *"An Overview of the Tesseract OCR Engine".*