2011 International Conference on Document Analysis and Recognition

Searching OCR'ed text: An LDA based Approach

Ehtesham Hassan, Vikram Garg, S. K. Mirajul Haque, Santanu Chaudhury, M Gopal
Department of Electrical Engineering
Indian Institute of Technology Delhi, New Delhi
{*hassan.ehtesham, gargvikram, miraj.bwn*}*@gmail.com,* {*santanuc, mgopal*}*@ee.iitd.ac.in*

***Abstract*—Indexing and retrieval performance over digitized document collection significantly depends on the performance of available Optical Character Recognition (OCR). The paper presents a novel document indexing framework which attends the document digitization errors in the indexing process to improve the overall retrieval accuracy. The proposed indexing framework is based on topic modeling using Latent Dirichlet Allocation (LDA). The OCR's confidence in correctly recognizing a symbol is propagated in topic learning process such that semantic grouping of word examples carefully distinguishes between commonly confusing words. We present a novel application of *Lucene* with topic modeling for document indexing application. The experimental evaluation of the proposed framework is presented on document collection belonging to Devanagari script.**

***Keywords*-Document Retrieval, Latent Dirichlet Allocation, Optical Character Recognition**

I. Introduction

The paper presents a novel topic modeling based document indexing framework for digitized document image collection. The objective of the present work is to define an indexing framework for OCR'ed documents which can accurately retrieve text documents, despite OCR's poor performance. The framework is defined without actually correcting the erroneous digitized text but exploiting performance characteristic of the OCR based on its character level confusion information.

The topic model based indexing groups different terms occurring in the text document based on their semantic relationship [1][2][3]. The learning process extracts the latent topics in the document space using the *tf-idf* (term frequency-inverse document frequency) based document representation. The *tf-idf* scheme represents each document by a vector representing the frequency of different 'terms' or 'words' in the document. The model explores the term-document space by modeling the occurrences of terms and documents using a mixture of topics. However, errors in the digitization process causes inaccurate estimation of term-document frequency leading to inaccurate retrieval results. The OCR technology for Indian scripts is not perfect, however large collection of digital documents is maintained by various digital libraries.

In this direction, the paper presents a novel indexing scheme for documents converted by imperfect OCR technology. The framework exploits the knowledge of OCR's confusion matrix such that the topic model captures the semantic relationship between words considering the confusing cases. For example, an OCR confusion between 'a' and 'o' may group 'capitol' and 'capital' together. The objective is to learn the topic model by alleviating the influence of recognition errors such that topic probability assignment incorporates the knowledge about confusing cases. In this case, the semantic grouping performed by topic model based search will be more robust and realistic. Using the learned topic model, the topic based document representation is further applied for indexing using *Lucene* [4].

The paper organization is as following: Section II present related works in the scope of present work. Section III presents proposed indexing and retrieval framework with brief discussion on Latent Dirichlet Allocation. The experimental results and related discussion is presented in section IV. Finally, section V concludes the presented work.

II. Related Works

Traditional word matching based indexing and retrieval schemes retrieve documents by comparing the query over the dataset examples. However, the semantic relevance of query terms with the retrieve documents is not validated. Additionally in the case of concept based query defined by multiple terms, word matching based retrieval schemes perform poorly. In such scenario, the retrieval framework should also include the semantic relation between query terms.

Topic model based indexing framework provide excellent solution for such application. In the context of document retrieval traditional vector space representation (*tf-idf*) does not represent any semantic information about the document as the terms are considered to be independent. Topic based stochastic modeling explores the latent topical structure of the documents by modeling semantically related terms to a topic. Therefore topic based document representation presents an intuitive approach for defining semantic retrieval schemes. Using the topic based representation, we can define a concept based retrieval scheme. Here a concept, is represented by semantic grouping of different terms, extracted by their frequency of occurrence over the document set. Topic models have been extensively applied for document summarization, and indexing applications [5][6][7][8][9]. However, topic based indexing and retrieval applications is

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.244

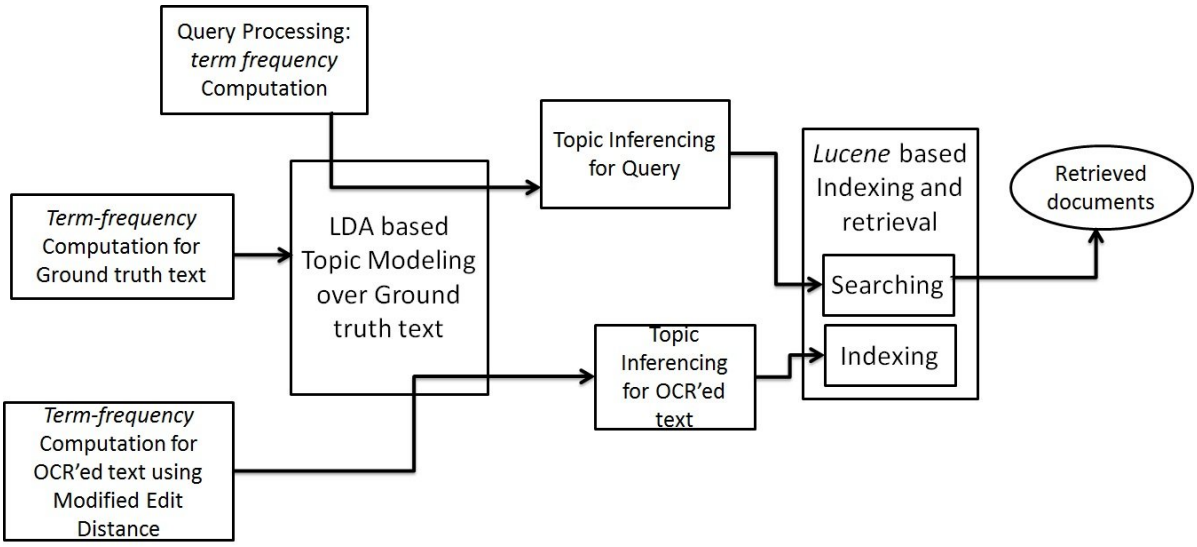

Figure 1. Proposed document indexing and retrieval framework

not explored much for document images. In this direction, some of the existing works have developed topic model based error correction framework as post-processing step of OCR [10][11].

However the error correction does not guarantee perfect OCR'ed output. The effect of recognition errors on retrieval using *tf-idf* based representation has been empirically studied by Taghva et al. [12]. The recent work by Walker et al. [13] have experimented the robustness of different topic models in the case of erroneous OCR'ed documents. In this direction, our work presents a novel framework to enhance the robustness of topic model based indexing of OCR'ed text without requiring explicit error correction after digitization. The approach exploits performance characteristics of the recognition for obtaining improved results. In particular, this technique has been used for document images of Devanagari for which the OCR technology is in under-developed stage. Additionally, the work presents novel application of LDA as very few attempts have been made in the past for its use in retrieving document images of Indian scripts. The proposed framework is tested on Devanagari script document collection prepared by low accuracy OCR which contains significant amount of digitization errors.

III. Document Indexing and Retrieval Framework

Proposed document indexing framework first learns the document topic model over a subset of ground truth data available for the OCR'ed document set. The figure 1 presents the flow diagram of the proposed indexing and retrieval framework. The learned topic model is further used to index the OCR'ed document by inferencing. The query document is initially inferred for getting the distribution of different topics over its terms. Corresponding to the query topic distribution we retrieve the relevant documents with respect to each topic. We have applied *Lucene* based indexing for retrieving the relevant documents corresponding to each document. The advantage of *Lucene* in combination with LDA is to provide fast search to retrieve documents corresponding to different topics. In the following discussion, we start with brief overview of LDA. The subsequent discussion presents the concept of modified edit distance for inclusion of the OCR's confusion characteristics for string similarity check. Section III-C and III-D presents the detailed discussion on proposed indexing and retrieval framework.

A. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) defines a generative probabilistic model over the document collection [3]. The documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The plate diagram for LDA generative process is shown in figure 2. The outer plate represents the documents and inner plate represents the topic sampling over set of words. M denotes number of documents in the collection and N represents a number of words in a document. The

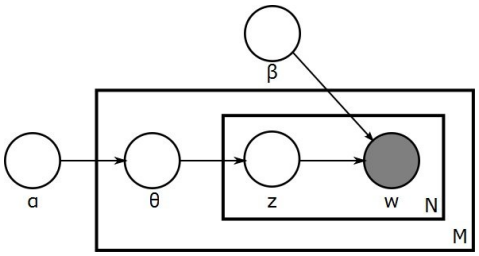

Figure 2. Graphical model for LDA

generative process of LDA for each document d in the collection D is defined as:

1) Draw topic distribution with Dirichlet prior:

$$\theta \sim \mathrm{Dir}(\alpha)$$

2) For each word w_n in d
   - Draw topic z_n from Multinomial over θ.
   - Draw word w_n from Multinomial conditioned over z_n.

Parameter β in the figure 2 represents topic/word probabilities as fixed quantities which needs to be estimated. However for most of the application smooth LDA model is applied which explicitly models β as random variable.

Inferencing

The LDA based inferencing requires the computation of posterior over the latent variables

$$p(\theta, z|d, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{1}$$

The distribution is unsolvable because of the intractable form of denominator because if the complex interdependencies between θ and β. The problem is solved by applying approximate inferencing based algorithms to obtain the approximate posterior estimation (e.g. Laplace approximation, variational approximation, and Markov chain Monte Carlo Simulation).

The estimation of model parameters α and $\beta_{1:K}$ requires maximization of marginal log likelihood of the data.

$$L(\alpha, \beta) = \sum_{i=1}^{M} p(w_d|\alpha, \beta) \tag{2}$$

As discussed the exact solution of likelihood function is not possible, however we can use approximate algorithms to obtain a tractable lower bound on the log-likelihood which can be maximized by expectation-maximization.

B. Inclusion of OCR confidence

The confusion matrix information for an OCR represents its confidence in recognizing various characters. The confusion matrix of the OCR can be learned using the document image collection and its ground truth data. The term-document vector representation in topic learning represents the frequency of different terms in the document. In the present case, we redefine the term-document matrix computation by defining modified edit distance based string matching. The modified edit distance is computed on the Levenshtein algorithm and incorporates the confusion matrix in the distance computation. The Levenshtein algorithm computes string distance as the minimum number of insertion, deletion and substitution operations required for transforming one string to another. The algorithm assigns uniform penalty for all such operations. In the present case, we redefine the substitution cost based on OCR's confusion between i^{th} and j^{th} characters as following.

$$\mathsf{Substitution\ Cost}(i, j) = 1 - \mathsf{OCR\_Conf\_mat}(i, j)$$

Once the modified edit distance between strings Str1 and Str2 is computed, the similarity of strings is concluded by applying threshold over the computed distance. The steps for string matching is described as following:

- Compute the modified edit distance between Str1 and Str2
- Normalize the distance by the length of longer string as the edit distance is upper bounded by length of longer string
- Conclude the string similarity by applying threshold over the normalized distance

The term-document vector for OCR'ed documents is generated using the above discussed approach for string similarity, which is subsequently used for topic inferencing.

C. Methodology for Indexing OCR'ed Documents

The section presents the listing of steps in our application of LDA based topic modeling, topic vectors based indexing using *Lucene* and document retrieval for the given query.

- For the complete document collection, use of the subset of ground truth data to prepare the vocabulary of unique terms.
- Preparation of the term-document matrix for ground truth data based on the occurrences of unique terms vocabulary.
- LDA model learning for semantically grouping the document terms existing in the ground truth.
- Preparation of the term-document matrix for OCR'ed documents using the vocabulary generated by ground truth. The term similarity is established by computing the modified edit distance as discussed in the section III-B. The selection of optimum threshold is based on the statistical analysis of the OCR output. We have varied the threshold within the range of average recognition error $\pm$ 10% of the variance of recognition accuracy. The selection of threshold parameter requires tuning such that a unique term in the vocabulary should not be concluded similar to highly dissimilar words.
- The term vector corresponding to each document is then converted into topic vector by inferencing the topic distribution using the learned LDA model.
- The OCR'ed document collection is further indexed using its topic distribution. The indexing is done using *Lucene*. The numeric field indexing scheme mechanism is used here instead of traditional term based indexing. Each topic is considered as a numeric field for a document, and topic-weight (probability) is considered its value. If we have k topic, the document d_i is converted into topic vector t_i as $\{t_{i1}, t_{i2}, ..., t_{ik}\}$. Each

t_{ij} is added as numeric field to corresponding *Lucene* document.

D. Retrieval framework

In the case of querying over a OCR'ed document converted by immature technology, correct retrieval is not always possible. However, it is expected that documents having similar topic distribution as ground truth should be correctly retrieved. The search query is converted into a topic vector i.e. $\{q = t_{i1}, t_{i2}, ..., t_{ik}\}$ by learned LDA model over ground truth. To retrieve the relevant documents, we use query topic vector to search for OCR'ed document's topic indices. The topic vector for q is searched in the *Lucene* index using numeric range query. The numeric range query compares the attributes within a range for scoring the documents. The numeric scores in this case, do not always help in distinguishing, or ranking the documents falling in the same range. Using a small range may fail to cover many relevant results, therefore achieving very low recall. Whereas larger range assigns same score to many documents, therefore reducing the precision of retrieval. Therefore, we used cosine similarity based score to rank the relevant documents returned by *Lucene* search. *Lucene* based search gives a faster retrieval performance cosine similarity is calculated for small set of documents returned by *Lucene*.

IV. Results and Discussion

In general, the OCR technology for Indian scripts is not matured yet, therefore we selected sample document image collection belonging to Devanagari script for experimental evaluation of the proposed indexing framework. The collection is prepared as part of consortium project funded by Government of India [14]. The digitization of the document collection is performed by OCR which achieves 77.8% accuracy at character level recognition. The collection contains 600 document images. We have evaluated the retrieval framework for 61 synthetic queries having 2 to 4 words.

The edit-distance threshold is selected as 0.1 for building the equivalent term set for a given term (refer section III-C for details). For performance measurement of the system, we have also indexed topic vectors for ground truth data using *Lucene*. We compute the overlap of retrieved documents for a query, over Ground truth text and OCR'ed text as the performance measure. The results are presented in the figure 3. The results establish the validity of the proposed framework. The overlap between the retrieved results improves significantly with increase in search radius. In information retrieval problem, it is always desired to have all the relevant documents in the top of ranking. In this context, the framework achieves encouraging results as the percentage of overlap is maximum in the case of Top-3 relevant results.

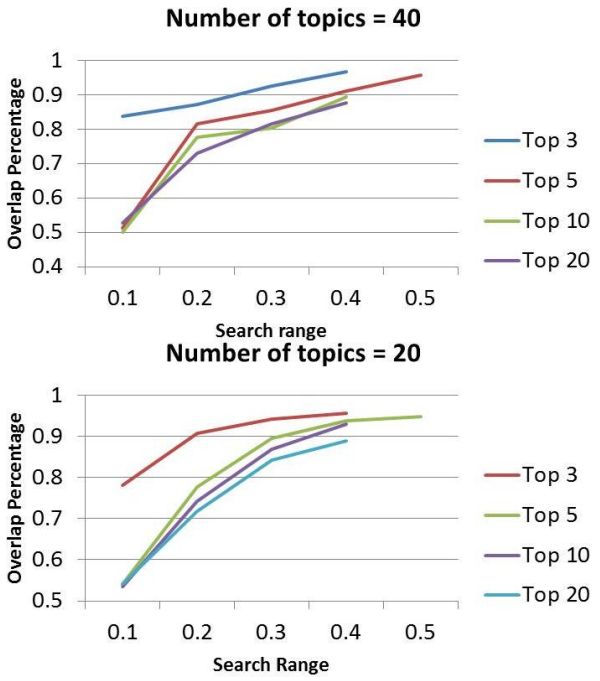

Figure 3. Percentage overlap for given query set over retrieved documents from ground truth and OCR'ed text

V. Conclusion

We have presented a novel document indexing and retrieval framework for semantic search over a text collection prepared by inaccurate OCR technology. The efficacy of the proposed framework is demonstrated over Devanagari script document collection. Our work presents novel application of *Lucene* for topic based search over the document collection. The evaluation of the proposed framework for other Indian scripts consists of the future works.

Acknowledgment

This work is funded by MCIT, Government of India as a part of project *Development of Robust Document Analysis and Recognition System for Printed Indian Scripts*.

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic indexing," *Journal of the Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[2] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, March 2003.

[4] [Online]. Available: http://lucene.apache.org/

[5] C. C. Aggarwal and P. S. Yu, "On effective conceptual indexing and similarity search in text data," *Proceeding of International Conference on Data Mining*, pp. 3–10, 2001.

[6] D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-document summarization using sentence-based topic models," *Proceeding of the ACL-IJCNLP 2009*, pp. 297–300, 2009.

[7] A. Takasu, "Cross-lingual keyword recommendation using latent topics," *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pp. 52–56, 2010.

[8] L. Zhang, J. Chang, X. Xiang, and X. Feng, "Topic indexing of spoken documents based on optimized n-best approach," *Proceedings of the International Conference on Intelligent Computing and Intelligent Systems*, 2009.

[9] E. H. Ramirez and R. F. Brena, "An information-theoretic approach for unsupervised topic mining in large text collections," *Proceedings of the International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pp. 331–334, 2009.

[10] F. Farooq, A. Bhardwaj, and V. Govindaraju, "Using topic models for ocr correction," *Interneational Journal of Document Analysis and Recognition*, vol. 12, pp. 153–164, October 2009.

[11] M. L. Wick, M. G. Ross, and E. G. Learned-Miller, "Context-sensitive error correction: Using topic models to improve ocr." *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 1168–1172, 2007.

[12] K. Taghva, J. Borsack, and A. Condit, "Effects of ocr errors on ranking and feedback using the vector space model," *International Journal of Information Processing and Management*, vol. 32, pp. 317–327, May 1996.

[13] D. D. Walker, W. B. Lund, and E. K. Ringger, "Evaluating models of latent document semantics in the presence of ocr errors," *Proceeding the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 240–250, 2010.

[14] [Online]. Available: http://ocr.cdacnoida.in/