RESEARCH ARTICLE OPEN ACCESS

Artificial Intelligence for Document Segregation

C.Navamani.MCA.,M.Phil.,ME[1], Naveendiran.K[2],
Assistant Professor[1], Final year[2]
Department of computer applications,
Nandha Engineering College/Anna University,Erode.

--**************************---------------------------------

Abstract:

The main concept of this project is to make machine to understand human language through NLP (Natural Language Processing).This NLP concept is based on the "ARTIFICIAL INTELLIGENCE". Artificial intelligence (AI) is the intelligence exhibited by machines or software. It is also the name of the academic field of study which studies how to create computers and computer software that are capable of intelligent behaviour.

Artificial intelligence for document segregation is used to Mapping the Lease id from PDF Content. We receive the bulk of lease documents, lease id and address from different clients then generate the file path, file name and size of the file in MS Excel document. After generate Excel file submit into the document segregation tool. This tool automatically segregates the documents by using each lease id and address.

Keywords — **Artificial intelligence, Natural language processing, Document segregation.**

--**************************---------------------------------

I. INTRODUCTION

This project helps the IT employees to reduce the manual work. It is also helps to reduce the time for manual segregation process. In manual process the employees can manually read the pdf content and finding the address from pdf content, after finding the address then compare to the client address if any one address matched means enter the lease id to the particular lease document. This process can be continued until all files are segregated.

This tool also used to some advance technologies (like Google map) for easily mapping the lease documents. In manual process various types of mistakes can be occurred, this tool provides the solution to avoid those problems and also used for mapping all documents successfully. Another one advantage of the tool is reducing the manpower for segregate the documents.

In order to search and retrieve the information efficiently in Document using Segregation tool, the client data set should be created for the documents with enough details. Various types of documents need for segregation process finished successfully. So we can use seven different steps for mapping the documents properly.

Presently, the effectiveness of ontology in document segregation scheme [3] has been realized as this technique is supposed to improve the accuracy of segregation. Whereas the role-played by ontology's [4], provides a position of legitimacy, research in this field leads to superior importance in the arising of different disputes featured in the modern digital situation. With the intention of providing solution to the various drawbacks related with present search techniques, ontology's are widely implemented for creating the capable document segregation techniques [5, 6].

II. RELATED WORK

Mapping of document is very important for the purpose of document segregation, summarization, topic extraction and information retrieval in an efficient way. Initially, segregation is applied for enhancing the precision or recall in the information retrieval techniques [1, 2]. In recent times, segregation technique is applicable in the areas which involves browsing a gathered data or in categorizing the outcome provided by the search engine for the reply to the query provided by the user. Document segregation can also be applicable in producing the hundred percentage of the result.

Proposed system of a mapping technique on segregation based topic extraction and classification. They have discussed on classification and extraction of news items in ePaper, a prototype system of a future personalized newspaper service on a mobile reading device. The ePaper system comprises news items from different news suppliers and distributes to each subscribed user a personalized electronic newspaper, making use of content-based and collaborative filtering techniques. The ePaper can also offer users standard version of chosen newspapers, besides the browsing abilities in the warehouse of news items. This deliberates on the automatic categorization of incoming news with the help of hierarchical news ontology. Based on this segregation technique on one hand, and on the client lease document details on the other hand, the personalization engine of the system is able to afford a personalized paper to every user onto the mobile reading device.

By considering the difficulty that classical Euclidean distance metric cannot create a suitable separation for data lying in a manifold, a GA based segregation method with the help of geodesic distance measure is proposed by Gang Li et al, [15]. In the proposed method, a prototype based genetic illustration is used, where every chromosome is a sequence of positive integer numbers that indicate the k-medoids. In addition, a geodesic distance based proximity measures is applied to find out the similarity between data points. Simulation results on eight standard synthetic datasets with dissimilar manifold structure illustrate the effectiveness of the algorithm as a segregation technique. Evaluating with generic k-means method for the function of segregation, the proposed technique has the potential to distinguish complicated non-convex documents and its mapping performance is obviously better than that of the K-means method for complex manifold structures.

Andreas et al, [13] discussed on the segregation technique for text data. Text mapping usually involves segregate in a high dimensional space that appears complex with considered to virtually all practical settings. Additionally, provided a scrupulous segregation outcome it is normally very tough to come up with a good clarification of why the text clusters have been created the way they are. In this paper, a mapping technique is presented for applying background information during pre-processing for improving the mapping outcome and permit for selection between outcomes. They have pre-processes the input data supplied to ontology-based heuristics for feature selection and feature aggregation.

Therefore, various choices for text illustrations are constructed. Based on these illustrations, they have calculated the multiple segregation outcomes using K-means. This put forth by Momin et al., [15]. Document mapping the achieved results by compared favourably with a sophisticated baseline pre-processing strategy.

Document segregation methods generally based on single term examination of document data set. To attain more precise document segregation, more informative feature like phrases are essential in this scenario. Therefore first part of the paper provides phrase-based model, Document Index Graph (DIG) that permits incremental phrase-based encoding of documents and capable phrase matching. It stress on efficiency of phrase-based similarity measure over conventional single term based similarities. In the second part, a Document Index Graph based Segregation (DIGBC) algorithm is provided to improve the DIG model for incremental and segregation. This technique incrementally clients documents based on presented final-document similarity measure. It permits assignment of a document to more than single document.

Muflikhah et al. [12] proposed a document mapping technique using concept space and cosine similarity measurement. This paper aims to incorporate the information retrieval technique and document segregation technique as concept space approach. The technique is known as Latent Semantic Index (LSI) approach which used Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA). The intention of this technique is to decrease the matrix dimension by identifying the pattern in document collection with refers to simultaneous of the terms. Every technique is employed to weight of term-document in vector space model (VSM) for document segregation with the help of fuzzy c-means technique. In addition to the reduction of term-document matrix, this research also utilizes the cosine similarity measurement as alternative of Euclidean distance to engage in fuzzy c-means.

Affinity-based similarity measure for Web document segregation is presented by Shyu et al., [5]. In this paper, the concept of document segregation is extended into Web document segregation by establishing the approach of affinity based similarity measure, which makes use of the user access patterns in finding the similarities among Web documents through a probabilistic model. Various experiments are conducted for evaluation with the help of real data set and the experimental results illustrated that the presented similarity measure outperforms the cosine coefficient and the Euclidean distance technique under various document segregation techniques.

ELdesoky et al., given a novel similarity measure for document segregation based on topic phrases. In the conventional vector space model (VSM) researchers have used the unique word that is contained in the document set as the candidate feature. Currently a latest trend which uses the phrase to be a more informative feature has considered; the issue that contributes in enhancing the document segregation accuracy and effectiveness. This paper presented a new technique for evaluating the similarity measure of the traditional VSM by considering the topic phrases of the document as the comprising terms for the VSM instead of the conventional term and applying the new technique to the Buckshot technique, which is a combination of the segregation technique and the K-means segregation method. Such a method may increase the effectiveness of the mapping by incrementing the evaluation metrics values.

Nanas et al. introduce a document evaluation function that allows the use of the concept hierarchy as a user profile for Information Filtering. Zitao et al., proposed a feature selection method for document segregation based on part-of-speech and word co-occurrence. Wang et al., presents a document mapping algorithm based on Nonnegative Matrix Factorization (NMF) and Support Vector Data Description (SVDD). Fuzzy mapping of text documents using Naive Bayesian Concept is provided by Roy et al., [12]. Web document segregation based on Global-Best Harmony Search, K-means, Frequent Term Sets and Bayesian Information Criterion is suggested by Cobos et al.,.

A document segregation method based on hierarchical algorithm with model segregation is presented by Haojun et al., This paper involved in analysing and making use of client technique to design client merging criterion. In this paper, they have presented a new method to calculate the overlap rate for improving time efficiency and the veracity. The technique is uses a line to pass across the two clients center as an alternative of the ridge curve. Depends on the hierarchical segregation technique, the expectation-maximization (EM) method is used in the Gaussian mixture model to count the parameters and formulate the two sub-clusters combined when their overlying is the biggest.

Document segregation with the help of fuzzy c-mean algorithm is proposed by Thaung et al., [11]. Most traditional segregation technique allocate each data to exactly single client, therefore creating a crisp separation of the provided data, but fuzzy mapping permits for degrees of membership, to which a data fit various clients. In this paper, documents are partitioned with the help of fuzzy c-means (FCM) mapping technique. Fuzzy c-means segregation is one of famous unsupervised mapping methods. But fuzzy c-means method needs the user to mention the number of clusters and different values of clusters corresponds to different fuzzy partitions earlier itself. So the validation of segregation result is required. PBM index and F-measure are helpful in validating the cluster. Yanjun et al., developed the feature selection algorithm for text segregation. Our research based on this research. For s improvement of textual features are represented and interact through ontology knowledge representation.

III. PROCESS FLOW OF DOCUMENT SEGREGATION TECHNIQUE

When we receive documents in bulk from a client the below mentioned approach will be explored to segregate the documents for each location. There are two major steps involved in segregating the documents such as follows:

1) Initial segregation using name of the file.
2) Inheriting Meaning Cloud, a text Analytics API.

The following inputs are required to do the document segregation

1) Document list:

It is an excel sheet which contains the list of "Document location path" and the "Document Name".

2) Inventory list:

It is an excel sheet which contains the list of Lease id and its Address information.

The following steps will be explored to achieve the Document segregation process.

Step 1: Including new columns in Document List. The Document list sheet will be updated with the Lease id, Street Address, City, Country column along with the existing columns "Document Location Path" and "Document Name".

Step 2: Mapping Document to Lease id based on Lease id in Inventory list

The fields Lease id, Address, City and Country from the Inventory list will be considered as reference fields for mapping to the appropriate document available in the Document list.

All the Lease id available in the Inventory list will be searched in the File Name column of the Document List. If matches, tool will update the corresponding Lease id in the Lease id column of Document list. Similarly the search operation will be performed by the tool using the address fields such as City and Country.

The tool will do the comparison of Street Address from the Inventory List to File Name of the Document List. If matches then the tool will update the appropriate lease id from the Inventory list to the Lease id column of the Document list. During the comparison by using the Street Address special characters and space will be ignored for comparison.

Step 3: Mapping Document to Lease id based on City and Country.

If there is only one record available with the particular city, then the tool will update the respective Lease id in the Document List.

Similar process will be followed to update the Lease id using the field Country also. fetching the address available in the inventory list in universal standard format using "Google Address Segregator".

Step 4: The address available in the Inventory list will be passed to the Google Address segregator to fetch the Addresses in a standard format. The Standard Street Address, Standard City, Standard Country, Latitude, Longitude address fields will be retrieved from the Google Address Segregator:

Step 5: Mapping Document to Lease id based on Standard Street Address. The Standard street address retrieved from Google Address Segregator/Address will be searched in the PDF document. If matches, the tool will update the respective lease id in the document List. This process will be done only for the PDF documents for which the lease id is not yet mapped. Other file extensions other than PDF will not be considered.

Step 6: Fetching all the addresses available in the Document that are listed in the Document List. This process will be done only for the PDF documents for which the lease id is not yet mapped. Meaning Cloud's "Topic Extraction" API will be consumed to scan the documents available in the Document list and to get all the address related information such as Street Address, City, State and Country available in each and every document.

The address retrieved from the Meaning Cloud Topic Extraction API will be passed to the Google Address segregator to fetch the Addresses in a standard format. The Standard Address, Standard City, Standard Country, Latitude, Longitude address fields will be retrieved from the Google Address Segregator:

Step 7: Updating the remaining Documents in Document list by appropriately mapping the Lease id. The Latitude and Longitude value retrieved from Google Address Segregator in Step 4(Inventory List) and Step 6(Document List) will be compared. There will be a tolerance level defined for this Latitude and Longitude comparison. The Lease id will be updated in the Document list for the Document records which has matching Latitude and Longitude.

If any documents found without Lease id needs to be analysed and segregated manually.

IV. CONCLUSIONS

The text mapping and document segregation to carry out the text document grouping. The text document contents were optimized with semantic analysis. The data mining domain based ontology is used to produce document segregation system. This system was found that document segregation appeared better than term mapping in terms of F-measure. In further work the term mapping to be converting into distributed term mapping and document segregation. Now the document segregation supports only English documents in future develop the tool for supporting various languages.

ACKNOWLEDGMENTS

The authors would like to thank M. Mitra for providing the WSJ data set and for his help with the SMART system, and the anonymous reviewers for several insightful comments.

N. L. Bhamidipati would like to thank the ISI-INSEAD (France) Fellowship for carrying out his doctoral research.

REFERENCES

[1] J. VanRijsbergen,Pp.“Information170-178,1997. Retrieval”, Buttersworth, London, second edition, 1989.

[2] G. Kowalski,Retrieval“InformationSystems–TheoryNamed Entity-Based Document Clustering", IEEE

[3] Momin, B.F., Kulkarni, P.J. and Chaudhari, A., "Web Document Clustering Using Document Index Graph", International Conference on Advanced Computing and Communications, Pp. 32 - 37, 2006.

[4] Muflikhah, L. and Baharudin, B., "Document Clustering Using Concept Space and Cosine Similarity Measurement", International Conference on Computer Technology and Development, Vol.1, Pp. 58-62, 2009.

[5] Shyu, M.L., Chen, S.C., Chen, M. and Rubin, S.H., "Affinity-based similarity measure for Web document clustering", IEEE International Conference on Information Reuse and Integration, Pp. 247 - 252, 2004.

[6] ELdesoky, A.E., Saleh, M. and Sakr, N.A., "Novel Similarity Measure for Document Clustering based on Topic Phrases", International Conference on Networking and Media Convergence, Pp. 92-96, 2009.

[7] N.Nanas, V.Uren and A. de Roeck, “Building and Applying a Concept Hierarchy Representation of a User Profile”, In Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 2003.archive/macros/latex/contrib/supported/IEEEtran/

[8] Zitao Liu, Wenchao Yu, Yalan Deng, Yongtao Wang and Zhiqi Bian, "A feature selection method for document clustering based on part-of-speech and word co-occurrence", Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Vol. 5, Pp. 2331-2334, 2010.

[9] Ziqiang Wang, Qingzhou Zhanga and Xia Sun, "Document clustering algorithm based on NMF and SVDD", Second International Conference on Communication Systems, Networks and Applications (ICCSNA), Vol. 1, Pp-192-195, 2010.

[10] Roy. R.S. and Toshniwal. D, "Fuzzy Clustering of Text Documents Using Naïve Bayesian Concept", International Conference on Recent Trends in Information, Telecommunication and Computing (ITC), Pp. 55-59, 2010

[11] Thaung Win. and Lin Mon., "Document segregation by fuzzy c-mean algorithm", 2nd International Conference on Advanced Computer Control (ICACC), Pp.239 - 242, 2010.

[12] Ding, Li et al. 2004. Swoogle: A Search and Metadata Engine for the Semantic Web. In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management. New York: ACM Press, pp. 652-659.

[13] Stekh Yu, Sardieh. F.M.E, Lobur. M and Dombrova. M, “Algorithm for clustering web documents”, Proceedings of VIth International Conference on Perspective Technologies and Methods in MEMS Design(MEMSTECH),Pp.187,2010.

[14] Lena Tenenboim., Bracha Shapira. and Peretz Shoval., "Ontology-Based Classification of News in an Electronic Newspaper", International Book Series Information Science and Computing, Pp: 89-98,2008.

[15] Gang Li., Jian Zhuang., Hongning Hou. and Dehong Yu., "A Genetic Algorithm based Clustering using Geodesic Distance Measure", IEEE International Conference on Intelligent Computing and Intelligent Systems,Pp:274–278,2009.