TELUGU LANGUAGE TEXT MINING

Y. Sri Lalitha

Associate Professor, Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology

ABSTRACT:

Text mining is crucial for extracting knowledge from important texts that are available in a variety of formats. These texts include pertinent information relating to the user's demand. In this research, we provide a tourist decision support system that extracts data from Telugu text files on tourist locations in both the Telugu-speaking states of Andhra Pradesh and Telangana, preprocesses the data, and divides the locations into three categories using the C 5.0 algorithm. The outcome is then put to use to assist foreign tourists in choosing points of interest that suit their preferences. The southern Indian states of Telangana and Andhra Pradesh both have Telugu as their official language. It is a language that more than 75 million people can read and write. Telugu language texts are used to retain information in a variety of formats, and the country has a rich cultural legacy. We also give a brief overview of our current and upcoming work employing field force automation and opinion mining approaches on the same tourist datasets.

Keywords— Text Mining, Decision Support System, Classification, C 5.0, machine translation.

[1] INTRODUCTION

Computer aided human decision making is greatly helped by the mining of pertinent data and the correct identification of rules and patterns in a huge database. In this study, we present the design and operation of a computer-assisted human decision-making system that aids foreign travellers in choosing emerging and uncharted tourist destinations. It is challenging for travellers to visit such interesting locations due to a dearth of information about them on the internet. Our method mines databases of local language tourist information to extract pertinent facts. This is possible because all the documents in the tourism corpus, more or less, follow the same written structure. The data extracted is then translated from the local language to English using a domain specific bi-lingual dictionary. Once translated into English, the data is then preprocessed for classification. The classifier is then trained and tested and the resulting decision tree or ruleset is used to classify unseen data. The classified data is then used to help tourists in selection of places to visit. This system has been designed for mining data in Telugu language. Telangana and Andhra Pradesh, two states in southern India with more than 4.2 billion readers and writers, both have Telugu as their official language. Telugu language texts are used to retain information in a variety of formats, and the country has a rich cultural legacy. However, the language lacks resources when it comes to natural language processing. WordNets, morphological analyzers, lemmatizers, stemmers, and other computational linguistic tools are few and few between, and they are often only used in academic research. As a result, Telugu is behind in applications linked to information retrieval and other relevant fields. Our main driving force behind creating the system is an intriguing one. The state depends heavily on tourism for its financial support, thus the administration is eager to encourage tourism in general and overseas travel in particular. The majority of the state's leading tourist attractions are well-liked by visitors from other countries. On the internet, there is a wealth of information on these places. However, the potential for these well-known tourist destinations to generate income has practically peaked. Information about the state's numerous emerging and uncharted places may be found in official papers. This paper have been digitised in Telugu as part of the bigger government digitization initiative. It is not practical to make these documents publicly available since they contain some sensitive information. These documents, which include a lot of irrelevant information, were initially not intended to be utilised in the tourist industry. Our key problem was to extract pertinent information from these documents and transform it into a form that may assist travellers in making decisions.

[2] LITERATURE COLLECTION

Since all documents are uniform in structure, we scan for certain cue words and phrases [1] to extract information.

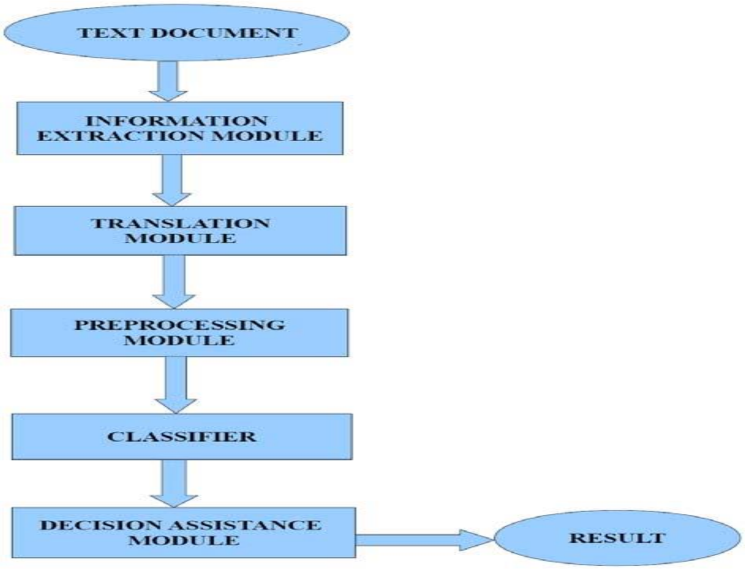

Fig. 1. Architecture of the proposed system

Presence of these cue words and phrases indicate that the relevant information is present. For example, the sentence containing information regarding distance of the place of interest from the nearest civilian airport will contain the Telugu phrase "*airport daggarlo aemi tourist places unnayi"* Similarly, to find the number of religious places in the area we scan for the Telugu phrase "*eeppudi mana dagaarlo gudi aemi vunnayi*" meaning that any worships or temple nearest to our place.

We convert the 26 properties from Telugu to English when they have all been extracted. One text, two yes/no type Boolean data, one alpha-numeric, and 22 numeric data make up the 26 properties. It is simple to translate numerical and Boolean data. Using common key mapping conventions for the Telugu-English language combination, the place's name has been transliterated. Once translated into English, the attributes are formatted as required by the C 5.0[2] classifier's input module. To divide the kind of the site of interest into one of three classes, we use the C 5.0 algorithm. A Type 1 tourist destination is a well-known, well-established tourist location with the most up-to-date amenities for visitors. Type 3 tourist destinations are undiscovered locations with minimal to no infrastructure and facilities, whereas Type 2 tourist destinations are developing locations with moderate facilities. Various classification algorithms have been used with varying results to classify textual data[2,3,4,5,6]. The ID3 method was succeeded by the C 4.5 algorithm, which is frequently employed in such jobs. However, it cannot accurately forecast for noisy data. Because of its sophisticated characteristics, we employ the C 5.0 classification algorithm. It offers boosting, a method for building and combining many classifiers for the same dataset. Boosting improves the predictability of the classifier. A lengthy decision tree could be challenging to read and understand. Another alternative offered by C 5.0 is to see the decision tree as a simple collection of rules. Additionally, C5.0 can foretell which traits will be important for categorization and which won't. Winnowing is a method that is particularly helpful when working with high dimensional datasets, or datasets with many properties. The fact that C 5.0 can manage missing data in the dataset is still another important justification for utilising it.

www.Jalletelugufadi.com
తెలుగు అక్షరమాల
అ ఆ ఇ ఈ ఉ ఊ ఋ ౠ ఎ ఏ ఐ ఒ ఓ ఔ
అం అః
క ఖ గ ఘ ఙ చ ఛ జ ఝ ఞ ట ఠ డ ఢ
ణ త థ ద ధ న ప ఫ బ భ మ య ర ల
వ శ ష స హ ళ క్ష ఱ

Telugu alphabets

The classifier is first evaluated using a test set and then a second, unclassified dataset where the class to be predicted is represented by?. The classifier is trained using the training set. The outcome is then saved for the decision-making module to utilise. A traveller can choose potential tourist places to visit with the aid of the decision-making module.

The tourist must specify the type of destination he is interested in visiting, i.e., whether he wants to go to a well-known tourist destination with the majority of facilities, whether he wants to go to a new destination with a moderate amount of facilities, or whether he wants to go to an uncharted area that offers a date with untouched resources but has few or no tourist facilities. Additionally, the visitor can indicate his preferred areas, such as environment and animals, water features, adventure spots, etc. The user question is subsequently processed by a straightforward query processor, which also gives relevant points of interest and the appropriate information. The user may order the results based on any single attribute or set of related qualities.

[3] RESULTS AND DISCUSSION

Utilizing the full tourist corpora managed by the state government, we tested our method. There have been 5140 classified papers used, or documents with a known kind of destination. Three thousand of such papers were utilised to train the classifier. The classifier was tested using the remaining 1140 documents. After that, 3000 new documents' type of location was predicted using the classifier. The classifier's accuracy is displayed in Table 2 With 2061 accurate predictions on the testing set, 96 percent accuracy was attained. The accuracy on the hidden dataset was 94%, with 3758 accurate predictions. The confusion matrices for the test and unseen datasets are displayed in tables 3 and 4. For the test dataset, three type 1 instances were categorised as type 2, and one example as type 3. 43 instances of type 2 were categorised as type 3, whereas 32 cases of type 1 were. Three type 3 instances were categorised as type 2. Three occurrences of type 1 were categorised as type 2 in the unseen set. Type 1 classification was given to 12 instances of type 2 and type 3 to 149 cases. 78 instances of type 3 were misclassified as type 2 in error.

DataSet	Accuracy	Total cases
Test	96	3000
Unseen	94	2061

Table 2 : Accuracy of the classifier

	Type 1	Type 2	Type 3
Type1		3	1
Type 2	32		43
Type 3	0	3	------

Table 3: Confusion matrix

	Type 1	Type 2	Type 3
Type1		3	0
Type 2	12		149
Type 3	0	78	-----

Table 4: Confusion matrix

[4] CONCLUSION

We have discussed the layout and operation of a system that aids travellers in making destination decisions based on their preferences, even for regions where there is few information online. On an unknown dataset, the system performs with a 94 percent accuracy rate. We had a significant issue where numerous papers included missing values. We also had to figure out how to handle source documents that were both Unicode compliant and not. The documents digitized earlier were encoded using ISCII (Indian Script Code for Information Interchange) fonts which are not Unicode compatible. We are currently developing an administrative decision support module using the same tourism dataset for the project's next phase, which will be able to forecast trends in funding distribution, revenue generation, visitor volume, etc. so that promising tourist destinations can be developed appropriately. When making such judgments, it is intended to take into account visitor comments at these locations. The visitor's comments will be transmitted to the main server via field force automation methods. Volunteers will take care of this duty at the already-existing tourist kiosks at the tourist hotspots. On the input that has been provided, opinion mining techniques will be used, and the findings will be integrated with the administrative decision support module.

References

[1] Cohen, W. W., & Singer, Y. Context-sensitive learning methods for text categorization.In SIGIR '96: Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 307-315. (1996).

[2] Cover, T. M., & Thomas, J. A.. Elements of Information Theory. John Wiley and Sons, New York. (1991) 12.Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery, S. Learning to extract symbolic knowledge from the World Wide Web. In Proceedings of the Fifteenth National Conference on Artificial Intellligence (AAAI-98), pp. 509-516. (1998).

[3] Dagan, I., & Engelson, S. P. Committee-based sampling for training probabilistic classifiers. In Machine Learning: Proceedings of the Twelfth International Conference (ICML '95), pp. 150-157. (1995).

[4] Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B,39 (1), 1-38. (1977).

[5] Dietterich,T.G. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10 (7). (1998).

[6] Domingos, P., & Pazzani, M.. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29, 103-130. (1997).

[7] Friedman, J. H.. On bias, variance, 0/1-loss, and the curse-ofdimensionality. Data Mining and Knowledge Discovery, 1 (1), 55-77. (1997).

[8] Lewis, D., and Ringuette, M., “A Comparison of Two Learning Algorithms for Text Categorization,” In Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93, (1994) Hassan,M.M., Rahman,.C.M. “Text Categorization Using Association Rule Based Decision Tree,” Proceedings of 6th International Conference on Computer and Information Technology, JU,pp. 453-456, (2003)

[9] McCallum, A., and Nigam, K., “A Comparison of Events Models for Naïve Bayes Text Classification,” Papers from the AAAI Workshop, pp. 41-48, (1998).

[10] Yang Y., Zhang J. and Kisiel B, “A scalability analysis of classifiers in text categorization,” ACM SIGIR'03,(2003).