ISSN:0975-9646

Oriya Language Text Mining Using C5.0 Algorithm

Sohag Sundar Nanda, Soumya Mishra, Sanghamitra Mohanty

P.G. Department of Computer Science and Application

Utkal University, India

sohag_sundar_nanda@hotmail.com

soumyalitun@gmail.com

sangham1@rediffmail.com

Abstract: Text Mining is essential for knowledge discovery from valuable texts available in many forms. These texts carry relevant information pertaining to the need of the user. In this paper we describe a tourist decision support system that mines data regarding tourist places in Orissa from Oriya text files, translates and preprocesses data and classifies the tourist places into three classes using C 5.0 algorithm. The result obtained is then used to help international tourists in selecting places of interest according to their choice. Oriya Language is the official language of Orissa, a state in the eastern part of India. More than 31 million people speak and write this language. It has a rich heritage and culture and knowledge is stored in many forms through Oriya language text. We also present a sketch of our ongoing and future work on the same tourism datasets using field force automation and opinion mining techniques.

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

I. INTRODUCTION

Mining relevant data and determination of accurate rules and patterns in a large database acts as an important aide in computer assisted human decision making. In this paper we describe the architecture and working of computer assisted human decision making system that helps international tourists select developing and unexplored places of interest. Due to lack of information regarding such places on the internet, it becomes difficult for tourists to visit such places of interest. Our system mines local language tourism information datasets and extracts necessary information. 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 Oriya language. Oriya Language is the official language of Orissa, a state in the eastern part of India having more than 4.2 billion readers and writers. It has a rich heritage and culture and knowledge is stored in many forms through Oriya language text. However, from a Natural Language Processing point of view, the language is resource poor. Computational linguistic resources such as WordNets, morphological analyzers, lemmatizers, stemmers etc are far and few and are generally restricted to academic research. As a result, Oriya lags behind behind in information retrieval and other related applications. Our primary motivation for designing the system is an interesting one. Since tourism plays an important role in the providing revenue to the state, the government is keen to promote tourism, especially international tourism. Most top tourist destinations of the state are popular with international tourists. A lot of information is available on the internet related to these destinations. However, revenue earning potential of these established places of interest has almost reached a peak. The state is endowed with a lot of emerging and unexplored destinations, information on which is available in government documents. As a part of the larger government digitization program, these documents have been digitized in the Oriya language. Since some sensitive data is included in these documents, making them available publicly is not feasible. These documents were originally not meant to be used in the tourism domain and contain a lot of non-relevant data. Our main challenge was to extract relevant data from these documents and translate them into English so that they could help in decision making for tourists.

This paper is divided into four sections. Section 2 describes the architecture and working of our system. In section 3 we present the results obtained by the classification module used in our system. Section 4 summarizes our work and also gives a sketch of our future work.

II. ARCHITECTURE AND WORKING

Figure 1 shows the architecture of our system. The first task at our hand is to extract relevant data from Oriya language documents. We select a total of 26 attributes from the documents namely :

document id, name of the place of interest, whether district headquarters, total number of tourist places in the place of interest as surveyed by the State Tourism Department, number of five star and higher hotels, number of four star hotels, number of three star hotels, number of two star hotels, number of one star hotels, number of unrated hotels, number of government guest houses open for tourists, whether the place of interest is connected by motorable road with major cities in the state, distance from nearest civilian airport in kilometers, distance from nearest railway station in kilometers, past fiscal year financial outlay in lakhs of Rupees for the place of

interest, past fiscal year revenue earned by the place of interest, number (in lakhs) of tourists who visited last year, number of wildlife sites (animal reserves, bird sanctuaries, endangered species areas etc) in the place of interest, number of religious sites (historical temples and other places of worship etc) in the place of interest, number of waterbody sites (beaches, springs, confluence points, waterfalls etc) in the place of interest, number of medical tourism sites in the place of interest, number of art and craft sites in the place of interest, number of adventure sites in the place of interest, number of supermarkets and shopping malls in the place of interest, number of cinema halls including multiplexes in the place of interest and type of the place of interest.

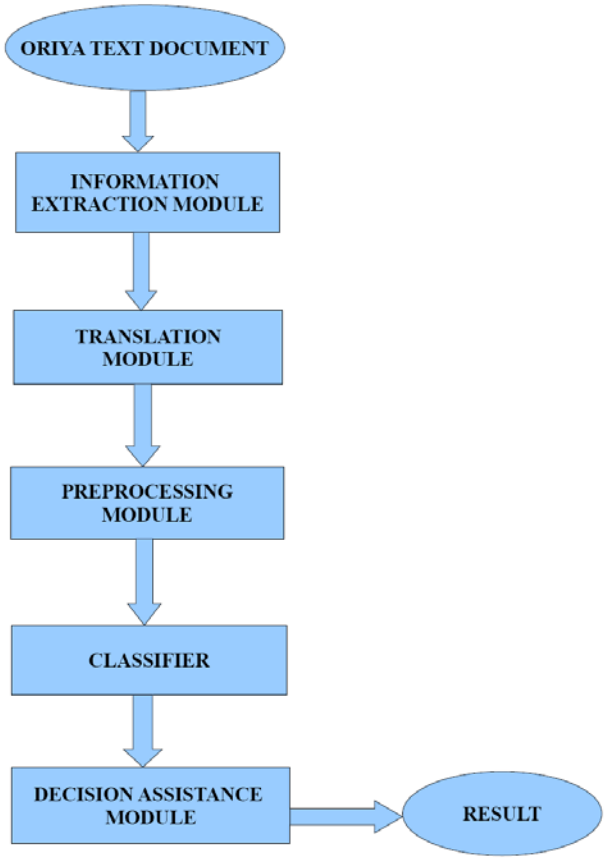

Fig. 1. Architecture of Tourist Decision Assistance

Since all documents are uniform in structure, we scan for certain cue words and phrases[1] to extract information. 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 Oriya phrase "*X nikatabarti bimanabandara tharu Y kilometer dura re*" meaning that the place X is Y kilometers away from the nearest airport. Similarly, to find the number of religious places in the area we scan for the Oriya phrase "*A re B ti aitihasika mandira o anyanya puja sthali achi*" meaning that the place A had B number of religious temples and other places of worship. Accordingly we extract Y and B from the former and the later sentences respectively. Any missing data for an attribute is represented by a question mark. Figure 2 shows the snapshot of an Oriya document used as input in the system while table 1 shows some of the cue words and phrases used by us.

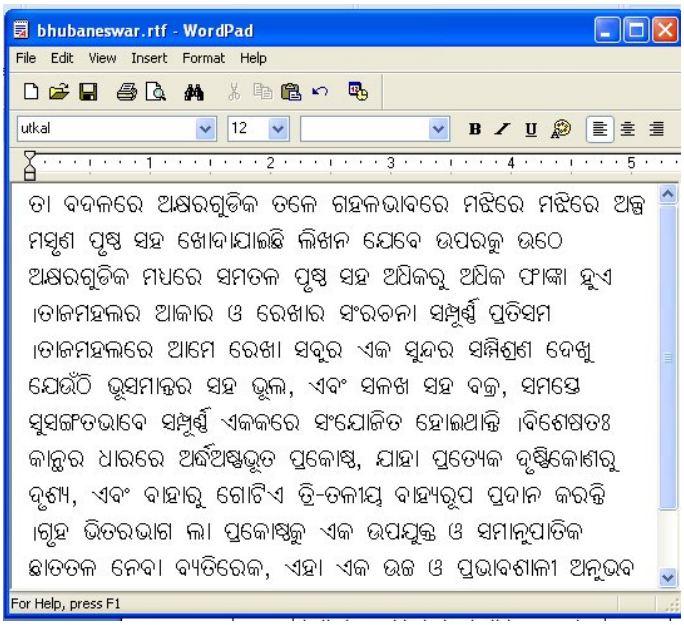

Fig. 2. Snapshot of input Oriya text

After extracting all 26 attributes, we translate them from Oriya to English. Out of the 26 attributes, 22 are numeric data, 2 are yes/no type Boolean data, one alpha-numeric and one (name of the place) is a string. Translation of numeric and Boolean data is obvious. The name of the place is transliterated using standard key mapping rules for Oriya-English language combination. Once translated into English, the attributes are formatted as required by the C 5.0[2] classifier's input module. We choose C 5.0 algorithm to classify the type of the place of interest into one of three classes. A Type1 tourist destination is a major and established tourist place with most modern facilities for the tourist. Type 2 tourist destinations are emerging destinations with moderate facilities while Type 3 tourist destinations are unexplored destinations with little or no infrastructure and facilities.

Various classification algorithms have been used with varying results to classify textual data[2,3,4,5,6]. The C 4.5 algorithm is a successor of the ID3 algorithm and is commonly used in such tasks. However it has limitations in predicting for noisy data. We use C 5.0 classification algorithm due to its advanced features. It provides for Boosting, a technique of construction and combination of multiple classifiers for the same dataset. Boosting increases the classifiers accuracy of prediction. A large decision tree may be difficult to read and comprehend. C 5.0 also provides the option of viewing the decision tree as a set of rules which is easy to understand. C 5.0 can also predict which attributes are relevant in classification and which are not. This technique, known as Winnowing is especially useful while dealing with high

dimensional datasets i.e. datasets with a large number of attributes. Another major reason for using C 5.0 is its ability to handle missing data in the dataset.

TABLE 1

Oriya Cue Words and Phrases

Oriya Cue Words/Phrases	English Translation
X re Y ti jalashaya achi	X has Y waterbodies
X ku Y koti tanka diya heichi	X has been granted Y crore rupees
X re zilla mukhyalaya rahichi	X is a district headquarter
X ku pacca rasta achi	X is connected by a motorable

The classifier is trained using the training set and then tested, first using a test set and further using an unseen and unclassified dataset where the class to be predicted is represented by ?. The result is then stored to be used by the decision making module.The decision making module helps a tourist in selecting prospective tourist destinations to visit.

The tourist needs to specify his type of destination i.e. whether he is interested in visiting a well established tourist destination with most facilities, or whether he wants to visit an emerging destination with moderate facilities or whether he is interested in visiting an unexplored location that offers a date with untouched resources but having no or little tourist facilities. Further, the visitor can also specify his area of liking like nature and wildlife, waterbodies, adventure sites etc. A simple query processor then processes the user query and provides relevant places of interest along with the details. The user can rank the results according to any specific attribute or combination of attributes. Figure 3 shows a snapshot of the decision assistance module.

III. RESULTS AND DISCUSSION

We have tested our system using the entire tourism corpora maintained by the state government. The total number of classified documents, i.e. with determined type of destination, used is 6143. Out of which 4000 documents were used for training the classifier. The rest 2143 documents were used for testing the classifier. The classifier was then used for predicting the type of location for 4000 new documents. Table 2 shows the accuracy of the classifier. On the testing set, the accuracy achieved was 96 percent with 2061 correct predictions. On the unseen dataset the accuracy achieved was 94 percent with 3758 correct predictions. Table 3 and table 4 show the confusion matrices for the test and unseen datasets. For the test dataset 3 cases of type 1 were classified as type 2 and 1 case was classified as type 3. 32 cases of type 2 were classified as type1 and 43 cases were classified as type 3. 3 cases of type 3 were classified as type 2. In the unseen set 3 cases of type 1 were classified as type2. 12 cases of type 2 were classified as type 1 and 149 cases as type 3. 78 cases of type 3 were wrongly classified as type 2. The training set, test set and unseen set contained 737, 341, 829 instances of missing data respectively.

Fig 3. Snapshot of tourist decision assistance module

TABLE 2

Results

DataSet	Accuracy	Total cases
Test	96	2143
Unseen	94	4000

TABLE 3

Confusion matrix for test set

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

TABLE 4
Confusion matrix for unseen set

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

IV. CONCLUSION AND FUTURE WORK

V.

We have presented the design and working of tourist decision assistance system that helps tourists in selecting places to visit based on their preference including locations on which very little data is available on the Internet. The system performs with an accuracy of 94 percent on unseen dataset. A major problem encountered by us was the presence of missing values in many documents. We also had to solve the problem of handling a mix of Unicode compatible and non-Unicode compatible source documents. The documents digitized earlier were encoded using ISCII (Indian Script Code for Information Interchange) fonts which are not Unicode compatible.

As the next stage of the project we are working on development of an administrative decision assistance module on the same tourism dataset which can predict patterns in fund allocation, revenue earned, number of visitors etc so that promising places of interest are developed accordingly. It is planned to include tourist feedback at these places while taking such decisions. Using field force automation techniques, the tourist's feedback will be sent to the central server. This task will be handled by volunteers at existing tourist kiosks in the places of interest. It is planned to use opinion mining techniques on the feedback received and integrate the results with the administrative decision assistance module.

References

1.Hovy, E.H. Automated Text Summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, pp. 583–598. Oxford University Press, Oxford (2005)

2.Quinlan, J. R. Data Mining Tools See5 and C5.0. http://www.rulequest.com/see5-info.html

3.Quinlan, J. R. C4.5: Programs For Machine Learning. Morgan Kauffman.(1993)

4.Quinlan, J. R. Discovering rules by induction from large collection of examples. In D. Michie (ed.), Expert Systems in the Micro Electronic Age. Edinburgh, UK: Edinburgh University Press.(1979)

5.Ekbal,A.,Bandyopadhyay,S.: Bengali Named Entity Recognition Using Classifier Combination. ICAPR 259-262 (2009)

6.Ramakrishnan.G, Chitrapura.K.P., Krishnapuram.R, Bhattacharyya.P: A model for handling approximate, noisy or incomplete labeling in text classification. ICML 681-688(2005)

7. Blum, A., & Mitchell, T. Combining labeled and unlabeled data with co-training.In Proceedings of the 11th Annual Conference on Computational Learning Theory, pp. 92-100. (1998).

8.Castelli, V., & Cover, T. M.. On the exponential value of labeled samples. Pattern Recognition Letters, 16 (1), 105-111. (1995)

9.Cheeseman, P., & Stutz, J. Bayesian classification (AutoClass): Theory and results. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.),Advances in Knowledge Discovery and Data Mining. MIT Press. (1996).

10.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).

11.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).

13.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).

14.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)

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

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

17.Friedman, J. H.. On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery, 1 (1), 55-77. (1997)

18.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)

19. 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)

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

21. Yang Y., Zhang J. and Kisiel B, "A scalability analysis of classifiers in text categorization," ACM SIGIR'03,(2003)