AUTOMATIC E-CONTENT GENERATION

E.Iniya Nehru, National Informatics Centre, Chennai
T.Mala, Anna University, Chennai.

ABSTRACT

Automatic E-content generation is an innovative idea in the field of Natural Language Processing. It aims on developing an intelligent tutoring system in Tamil language. This system focuses on delivering personalized content in Tamil language to an individual user needs based on their learning abilities and interests. The system is divided into two parts. The first part involves the domain corpus collection and construction of the knowledge base. The knowledge base is constructed using text categorization and information extraction techniques. Categorization is responsible for classifying given Tamil documents to their target class. This is performed using the Naive Bayes (NB) algorithm. Information extraction system was constructed by developing information-extraction rules by encoding patterns (e.g. regular expressions) that reliably identify the desired entities or relations. The second part involves identifying the interests of specific user's and then user profiling the identified interests by looking through their browsing history. The topic analyzer is constructed to analyze the user's profile and evaluate the user's knowledge using intelligent evaluator system. Then the personalized content will be generated based on the knowledge level of the user.

This paper is organized as follows. Section 1 discusses the literature survey in the area of content generation, section 2 talks about the overall system architecture, section 3 talks about the categorization, section 4 gives details on information extraction system and section 5 gives the performance evaluation and conclusion of the proposed system.

1. LITERATURE SURVEY

The bulk of the text categorization work has been devoted to cope with automatic categorization of English and Latin character documents. El-Kourdi et al. used Naive Bayes algorithm to automatically classify non-vocalized Arabic web documents to one of five pre-defined categories [2]. Taeho Jo and Dongho Cho proposed an alternative approach to machine learning based approaches for categorizing online news articles [6].

Chinglai Hor et al. suggest the use of a hybrid RS-GA method to process and extract implicit knowledge from operational data derived from relays and circuit breakers [1]. S.Iiritano and M.Ruffolo described a prototype of a vertical corporate portal that implements a KDD process for knowledge extraction from unstructured data contained in textual documents [4]. Jose M.Alonso et al. presented a user-friendly portable tool designed and developed in order to make easier knowledge extraction and representation for fuzzy logic based systems [5]. Harith Alani et al. developed a tool called Artequakt. Artequakt automatically extracts knowledge about artists from the Web, populates a knowledge base, and uses it to generate personalized biographies [3].

2. OVERALL SYSTEM ARCHITECTURE

The overall system architecture as shown in the figure 2.1 explains the various functionalities required to perform the automatic content generation. Automatic content generation focuses on delivering personalized content in Tamil language to an individual user needs based on their learning abilities and interests. Knowledge base is constructed using document categorization and tourism related information extraction. This phase uses Naive bayes classification algorithm to classify the tourism documents as temples, hill stations and hotels. After the classification process, information should be extracted from those documents. The extracted information should fill the template of each document of a particular category. Thus the knowledge base was constructed using text categorization and information extraction.

An individual user's interests are identified and recorded to create a user profile. A user profile is specific to a user and is subjected to change over time. The Topic categorizer is used to categorize the topic based on user's query. The topic analyzer is used to analyze the user's profile and evaluate the user's knowledge using intelligent evaluator system. Based on the user's knowledge, intelligent evaluator system makes a decision to suggest the same topic or to suggest a new topic retrieved from the knowledge base. Then the personalized content will be generated based on the knowledge level of the user.

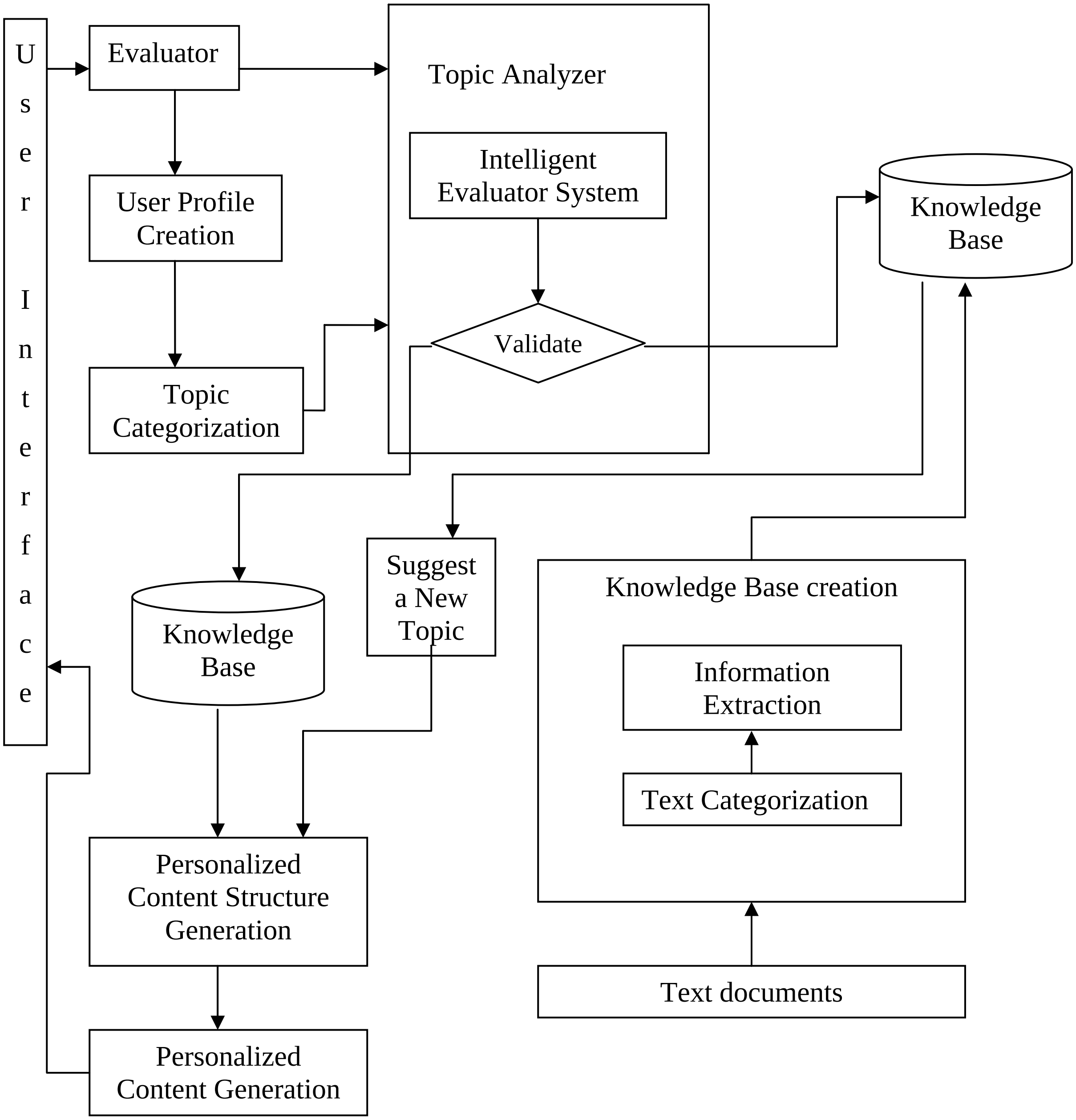

Figure 2.1 Detailed block diagram

3. DOCUMENT CATEGORIZATION

Document categorization is the task of automatically sorting a set of documents into categories from a predefined set. In the training phase, a set of documents are given with class labels attached, and a classification system is built using a learning method. Once the categorization scheme is learned, it can be used for classifying future documents. The following section gives a detailed view of preprocessing and classification of documents. A data preprocessing phase is required to weed out those words that are of no interest in building the classifier and also to reduce the processing time. Preprocessing phase comprises of two stages namely stop word removal and term pruning. In order to remove the stop words from the document, stop word list is prepared and the input tourism document is compared with the list and then stop words are removed. After removing the stop words, each document must be transformed in to a feature vector. This text representation, referred to as the bag-of-words. The problem of finding a "good" subset of features is called feature selection. Feature selection can be done by term pruning method. Term pruning is done according to the term frequency values. Here feature selection refers to relevant words. This would eliminate the very rare words that occurred and also the very common words that occur in almost every text document. Relevant words are then passed to the classification phase.

This phase uses naive bayes classification algorithm to classify the tourism documents as temples, hill stations and hotels. The NB classifier computes a posteriori probabilities of classes, using estimates obtained from a training set of labeled documents. When an unlabeled document is presented, the posteriori probability is computed for each class and the unlabeled document is then assigned to the class with the largest a posteriori probability.

4. INFORMATION EXTRACTION

Information extraction (IE) distills structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities. The document is tagged to identify the location, name of the place. Identify the Domain Events using rules i.e., retrieval of sentences based on heuristic rules. The heuristics rules are formed manually by analyzing the documents. Extraction of syntactic patterns from the sentences is done using part of speech information. Compare the syntactic patterns from the sentences with the predefined pattern. If it matches, then the corresponding template should be filled. The templates for various domain events are filled and then they are merged to form the complete filled template for the document.

5. PERFORMANCE EVALUATION AND CONCLUSION

From the 50 Tamil Nadu tourism documents, the first 30 are used for training and the next 20 are used for testing. The classification task considered here is to assign the documents to one the two categories. The Precision/Recall is used as a measure of performance for document categorization. Recall is the percentage of total documents for the given topic that are correctly classified. Precision is the percentage of predicted documents for the given topic that are correctly classified. The recall and the precision measures are calculated for each category and the average values are shown in the following table 5.1.

Table 5.1 Performance Evaluation Results Table

Category	Average Precision	Average Recall	Average F-measure
Kovil	100.00	80.00	88.89
Malai	83.33	100.00	90.71
Macroaverage	91.67	90.00	**89.80**

The result shows that the Naïve bayes achieves a macro average of 89.8. Thus it shows that the naïve bayes classifier performs well and its effectiveness is better in classifying the Tamil documents. The current work is confined to tourism documents of Tamilnadu state, but it can be extended to any category of documents. Categorization work can be further enhanced by adding more categories.

REFERENCES

1. Chinglai Hor, Peter A. Crossley, Dean L. Millar "Application of Genetic Algorithm and Rough Set Theory for Knowledge Extraction" IEEE transactions on Power Tech, pp. 1117 - 1122 , July 2007.

2. El-Kourdi M., Bensaid A. and Rachidi T. "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm". Proceedings of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages, pp. 51-58, August 2004.

3. Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal, Wendy Hall, Paul H. Lewis, and Nigel R. Shadbolt "Automatic Ontology-Based Knowledge Extraction from Web Documents" IEEE transactions on Intelligent systems, Vol.18, Issue no.1, pp-14-21, January 2003.

4. S.Iiritano M.Ruffolo "Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining" Database and Expert Systems Applications, 2001. Proceedings of 12th International Workshop on 3-7 September 2001, pp. 454 – 458, September 2001.

5. Jose M. Alonso, Luis Magdalena, Serge Guillaume "KBCT: A Knowledge Extraction and Representation Tool for Fuzzy Logic Based Systems" Proceedings of IEEE International Conference on Fuzzy Systems,Vol. 2, pp. 989-994, July 2004.

6. Taeho Jo and Dongho Cho, "Index Based Approach for Text Categorization", International journal of mathematics and computers in simulation, Issue 2, Volume 1, 2007.