International Journal of Engineering Inventions
ISSN: 2278-7461, ISBN: 2319-6491
Volume 2, Issue 1 (January 2013) PP: 31-36

Equirs: Explicitly Query Understanding Information Retrieval System Based on Hmm

Dilip Kirar[1], Pranita Jain[2]
[1]Research Scholar (M. Tech Student), Department of IT,
[2]Asst. Prof., Department of IT, Samrat Ashok Technological Institute, vidisha (M.P.)

Abstract:- Despite all the hypes, there are number of efforts has been taken to research in the field of Natural language processing but it has number of problems, such as ambiguity, limited coverage and lack of relative importance or we can say less accuracy in terms of processing. To reduce these problems and increase the accuracy we use "EQUIRS: Explicitly query understanding information retrieval system based on HMM". In this frame work, we use Hidden Markov Model (HMM) to improve the Accuracy and results, resolve the problem of ambiguity efficiently. Previously, various model used to improve the accuracy of text query, in which one of the most selective method is Fuzzy clustering method, but it is fail to reduce limited coverage problem. To reducing such problem and improving accuracy EQUIRS based on HMM and compare it with the result of fuzzy clustering techniques.

In the proposed frame work first 900 file is used to train which is divided into five file class categories called five query view cluster (organization, topic, exchange, place, people). Now, HMM is simply finding the nearest probability distance with the fired text query using QPU (Query Process Unit) and HMM will return suggestion based on emission probability (suggestion depth 5) which similar to query view. Thus proposed approach is different and has satisfied qualitative proficiency with using taxonomy of clustering (Precision, Recall, F-Measure, Training Time and Searching Time) from fuzzy based learning which has less accuracy

Keywords:-Information retrieval, Hidden Markov model, fuzzy cluster model, index.

I. INTRODUCTION

Natural language processing is becoming one of the most active areas in Human-computer Interaction. The goal of NLP is to enable communication between people and computers without resorting to memorization of complex commands and procedures. In other words, NLP is techniques which can make the computer understand the languages naturally used by humans. While natural language may be the Easiest symbol system for people to learn and use, it has proved to be the hardest for a computer to master. Despite the challenges, natural language processing is widely regarded as a promising and critically important endeavor in the field of computer research. The general goal for most computational linguists is to instill the computer with the ability to understand and generate natural language so that eventually people can address their computers through text as though they were addressing another person. The applications that will be possible when NLP capabilities are fully realized are impressive computers would be able to process natural language, translating languages accurately and in real time, or extracting and summarizing information from a variety of data sources, depending on the users' requests.

A hidden Markov model (HMM) [12] is a statisticalMarkov model in which the system being modeled is assumed to be a Markov process with unobserved (*hidden*) states. An HMM can be considered as the simplest dynamic Bayesian network.

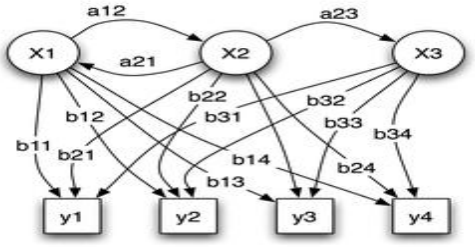


Figure1.1:- Hidden Markov Model

Probabilistic parameters of a HMM (example).
x — states
y — possible observations
a — state transition probabilities
b — output probabilities

In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a *hidden* Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; even if the model parameters are known exactly, the model is still 'hidden'

In this paper, we have evaluated training time, searching time and accuracy of the proposed algorithm. To measure these performance parameters we have used transaction data set that contains five file classes which is taken from Reuters-21578 text categorization test collection Distribution 1.0 README file [13].

As experimental result, the proposed algorithm retrieves information from large dataset with more training time and more searching time and also with great accuracy. The main purpose of the proposed algorithm is to improve precision, recall and accuracy.

II. BACKGROUND

Hidden Markov Models (HMM) [1] can also be used for classifying patterns from an unknown dataset. For example, in speech related literature HMM has been used for classifying speakers [2-3] or speech patterns [4, 5]. Typically, for pattern classification, a number of HMM are used in combination with supervised techniques. In this paper, we propose an EQUIRS based on HMM algorithm. In our model, a single HMM is used to identify the number of sequence and stats in a given dataset. The data items are then labeled and partitioned into the appropriate five file indexes. Initially, the HMM is used to calculate emission probability for each of the data items. Here, the emission probability on one hand represent how well the data fits the trained HMM and on the other provide a similarity measure between data items. While Hidden Markov Models have not been employed in web query classification, they have been extensively studied and applied in document classification [9], text categorization of multi-page documents [11], recognizing facial expressions from video sequences [8], and the infamous HMM part of speech tagger [7] and speech recognition [10]. While Cohen et al used the temporal facial expressions as the HMM states, speech recognition involves the phone symbols as the observation sequence [10]. Hidden Markov Models (HMM) were first introduced in the 1970s as a tool for speech recognition [6]. Recently, the popularity of HMM has increased in the pattern recognition domain primarily because of its strong mathematical basis and the ability to adapt to unknown data. This section describes HMM in more detail together with a description of the algorithms used to induce HMM. Further details can be found in [1].

The Hidden Markov Model (HMM) is a variant of a *finite state machine* having a set of hidden *states*, $\boldsymbol{Q}$, an output *alphabet* (observations), $\boldsymbol{O}$, transition probabilities, $\boldsymbol{A}$, output (emission) probabilities, $\boldsymbol{B}$, and initial state probabilities, $\boldsymbol{\Pi}$. The current state is not observable. Instead, each state produces an output with a certain probability ($\boldsymbol{B}$). Usually the states, $\boldsymbol{Q}$, and outputs, $\boldsymbol{O}$, are understood, so an HMM is said to be a triple, ($\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{\Pi}$).

Mathematical Definition:

Hidden states $\boldsymbol{Q} = \{ q_i \}$, $i = 1, \ldots, N$.

Transition probabilities $\boldsymbol{A} = \{a_{ij} = P(q_j \text{ at } t+1 \mid q_i \text{ at } t)\}$, where $P(a \mid b)$ is the conditional probability of a given b, $t = 1, \ldots, T$ is time, and q_i in $\boldsymbol{Q}$. Informally, $\boldsymbol{A}$ is the probability that the next state is q_j given that the current state is q_i.

Observations (symbols) $\boldsymbol{O} = \{ o_k \}$, $k = 1, \ldots, M$.

Emission probabilities $\boldsymbol{B} = \{ b_{ik} = b_i(o_k) = P(o_k \mid q_i) \}$, where o_k in $\boldsymbol{O}$. Informally, $\boldsymbol{B}$ is the probability that the output is o_k given that the current state is q_i.

Initial state probabilities $\boldsymbol{\Pi} = \{p_i = P(q_i \text{ at } t = 1)\}$.

III. PROPOSED WORK

In this paper, we propose a new model for natural language processing for text query information retrieval system is called EQUIRS: Explicitly Query Understanding Information Retrieval System based on HMM. These methods have significant theoretical advantages and it has shown impressive performance in many tasks such as text categorization test collection database, goal of text query understanding and automatic retrieve information probabilistic base is to compare the input text query vector with all the classes and then declare a decision that identifies to whom the input text query vector belongs to or if it doesn't belong to the database at

all. In this research work, Text query understanding is studied as an ambiguity and lack of knowledge problem. To tackle this problems problem our proposed model are considered in research work.

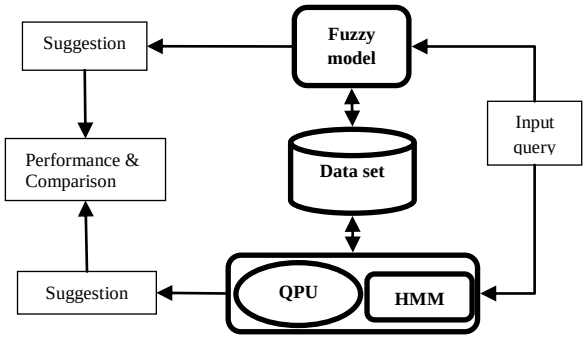


Figure1.2: - EQUIRS based on HMM architecture

Proposed Algorithm
Input- Training data set D (N is the total number of file; n is total number of training file), Input_query(1 to x), min_sug (suggession_depth)
Output - Performance and Comparison.

Training
Step 1:- Read all file data sets D (1 to N).
Step 2:- All data file N are convert it into class vector matrix and save in matrix vectors.
Step 3:- Apply it into hidden markov model in step 2.
Step 4:- After step 3 we calculate the emission probability Matrix (EMIS).
Step 5:- Store EMIS and Vectors.

Testing
Step 1:- Read input query (length 1 to x).
Step 2:- Input query are convert it into vector.
Step 3:- Load Transmission vector.
Step 4:- Add vector with training vector.
Step 5:- Hidden markov model are calculate Emission probabilities matrix.
Step 6:- Measure the most similar entries in step 5.
Step 7:- Calculate the similar entries vector in EMIS matrix.
Step 8:- Convert vector into string and display
Step 9:- End.

IV. RESULT

Performance Parameters
We measure the performance of our algorithm in the form of following parameters:
Training Time

Training time can be defined as the total time requires training the algorithm. There we generally compare the training time with fuzzy cluster model and EQUIRS based on HMM. In the previous fuzzy based model k-means algorithm is used divide knowledge into cluster due to this is required less time to training i.e. $\log_2 n$. Where as in our proposed approach will have to take more time to training then fuzzy because is use HMM. In which may sequence of state is generated. Which will take approx. $O(\log_2 n)$ time to train.

Searching Time

The searching time can be defined as total amount of time required to fining or retrieving a result. Generally it is important for any algorithm for its efficiency and always tries to keep minimum. However, in over algorithm is take more searching time then fuzzy based model roughly our model take O (log2n) time approximately which is equivalent to complexion of binary search.For classification tasks, [15] the terms **true positives**, **true negatives**, **false positives**, and **false negatives** compare the results of the classifier under test with trusted external judgments. The terms *positive* and *negative* refer to the classifier's prediction (sometimes known as the *expectation*), and the terms *true* and *false* refer to whether that prediction corresponds to the external judgment (sometimes known as the *observation*)

Precision

In our proposed approach EQUIRS based on HMM calculate the precision of information retrieval, precision is the fraction of retrieved documents that are relevant to the search:

$$\text{Precision} = \frac{tp}{tp + fp}$$

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n.

Recall

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Recall} = \frac{tp}{tp + fn}$$

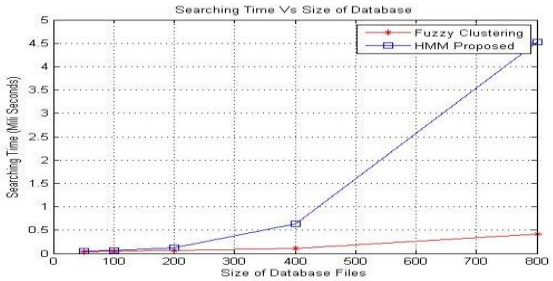
For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned

F_measure

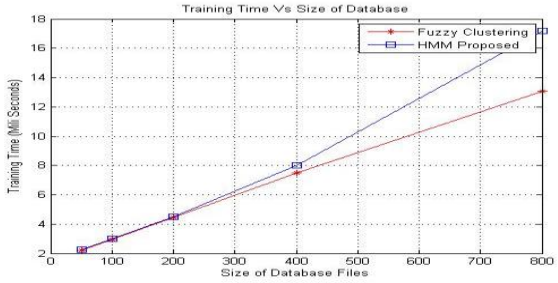
A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

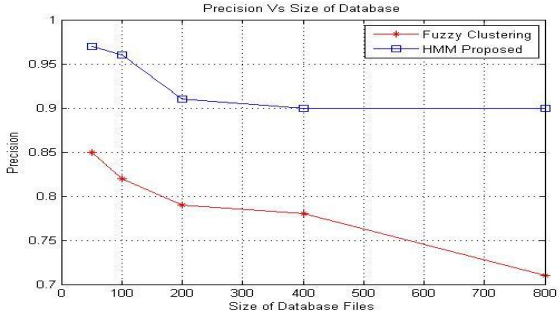
This is also known as the F_measure, because recall and precision are evenly weighted.



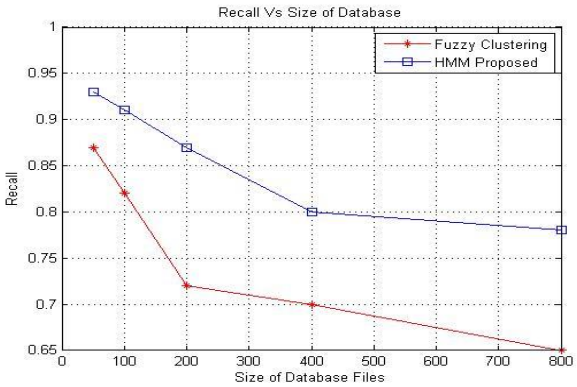
Graph1.1: Graph shows the training time difference between previous Fuzzy clustering approaches and our proposed approach EQUIRS based on HMM.



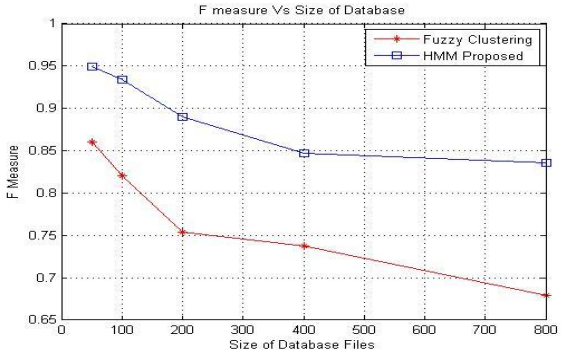
Graph1.2: Graph shows the searching time used difference between previous fuzzy clustering approach and our proposed approach EQUIRS based on HMM.



Graph1.3: Graph shows the precision difference between previous fuzzy clustering approach and our proposed approach EQUIRS based on HMM.



Graph 1.4: Graph shows the recall difference between previous fuzzy clustering approach and our proposed approach EQUIRS based on HMM.



Graph1.5: Graph shows the F_Measure difference between previous fuzzy clustering approach and our proposed approach EQUIRS based on HMM.

The above graph shows the result comparison which is generated by our proposed approach (EQUIRS: Explicitly Query Understanding Information Retrieval System based on HMM) and the previous method which is based on FCM [14]. In every graph it is clear that the time of training time, searching time, precision, recall and F_measure. Training time and searching time more than previous approach. The blue line is our EQUIRS approach based on HMM that takes more time to compute the result and red line indicate the fuzzy cluster model approach which takes less time in result generation for training time and searching time. Due the algorithm our proposed approach EQUIRS: Explicitly Query Understanding Information Retrieval System based on HMMis also taking less memory because HMM calculate the emission probability on current state not previous state. The graph 1.3, 1.4 and graph 1.5 gives the clear indication of the (94%) efficient accuracy usage of previous fuzzy clustering model approach and the proposed EQUIRS: Explicitly Query Understanding Information Retrieval System based on HMM approach.

Input	Fuzzy Cluster Model(FCM)					EQUIRS based on HMM Model				
	p	R	F	TT	ST	P	R	F	TT	ST
50	0.85	0.87	0.8598	2.2214	0.0367	0.97	0.93	0.9495	2.2525	0.0525
100	0.82	0.82	0.82	2.9532	0.0483	0.96	0.91	0.9343	3.0089	0.0705
200	0.79	0.72	0.5956	4.4422	0.0671	0.91	0.87	0.8895	4.4835	0.1243
400	0.78	0.70	0.7378	7.4741	0.1100	0.90	0.80	0.8470	8.0227	0.6365
800	o.71	0.65	0.6786	13.0528	0.422	0.90	0.78	0.8357	17.1637	4.5247

Table 1:- Table show result of both FCM and EQUIRS based on HMM proposed approach when taking different query and generate result.

The table 1 show the result comparison of fuzzy cluster match (FCM) and hidden markov model (HMM) when generating the result in term of precision (P), recall (R), F_measure (F), training time (TT) and searching time (ST).In this we take different text query and generate the result with both the approaches and

stored the precision, recall, F_measure, training time and searching time. As show in table for different number of text query for parameter training time and searching time of our proposed approach EQUIRS is greater than previous approach FCM. Therefore our proposed approach which is based on HMM is more efficient accuracy than the FCM approach.

V. FUTURE WORK

In this work we proposed an information retrieval system. We use emission probabilities based on likelihood sequence of state based Hidden Markov Model (HMM). Experiments on textual queries in multiple domain show that the proposed approach can improve the performance of which using taxonomy of clustering (Precision, Recall, F_measure, Training time and Searching time) significantly.

Know potentially the same approached can be applied to spoken queries given reliable speech recognition. For future work we will applied the lexicon modeling approached to larger datasets , we will also explore the use of other external resources such as Wikipedia for automatic learning.

REFERENCE

1) Rabiner R. L. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proceedings of the IEEE, Vol 77 (2), pp 257286, 1989.
2) Sadaoki F. *Speaker Recognition*. http://cslu.cse.ogi.edu/HLTsurvey/ch1node9.html
3) Ajmera J., Bourlard H., Lapidot I. and McCowan I. *Unknownmultiple speaker clustering using HMM,* InternationalConference on Spoken Language Processing, pp. 573576, 2002.
4) Huang X., Ariki Y. and Jack M. *Hidden Markov Models for speech recognition,* Edinburgh University Press, 1990.
5) Xie H., Anreae P., Zhang M. and Warren P. *Learning models for English speech recognition.* Proceedings of the 27th Conference on Australasian Computer Science, pp 323329, 2004.
6) Hassan M. R. and Nath B. *Stock Market Forecasting using Hidden Markov Model: A new approach.* Proceedings of International Conference on Intelligent Systems Design and Applications, IEEE Computer Society Press, pp. 192196, 2005.
7) Bernard Merialdo, 1994. "Tagging English text with probabilistic model", *Computational Linguistics 20*, pp. 155–171.
8) Cohen, A.M.I., Sebe, N. and Huang, T.S., 2002. "Facial expression recognition from video sequences", *In Proceedings of IEEE International Conference on International Conference on Multimidia& Expo*, pp. 121 – 124.
9) NikolaosTsimboukakis, George Tambouratzis, 2008. "Document classification system basedon HMM word map, *In Proceedings of CSTST*', pp. 7-12.
10) Lawrence R. Rabiner, 1989. "A tutorial on Hidden Markov Models and selected applications inspeech recognition", *Proceedings of the IEEE 77 (2),* pp. 257–286.
11) Paolo Frasconi, Giovanni Soda, Alessandro Vullo, 2002. "Hidden Markov Models for TextCategorization in Multi-Page Documents", *Journal of Intelligent Information Systems* 18:2/3,pp. 195–217.
12) http://www.nist.gov/dads/HTML/hiddenMarkovModel.html:
13) http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection.
14) *Jingling Liu1, Xiao Li2, Alex Acero2 and Ye-Yi Wang2, 2011* LEXICON MODELING FOR QUERY UNDERSTANDING. In proceedingsofICASSP 2011.
15) http://en.wikipedia.org/wiki/Precision and recall - Wikipedia, the free encyclopedia.