A Survey on: Automatic Question Tagging and Student Performance Prediction System.

Ms. Shital Kakad[1], Shivani Gautam[2], Swati Suman[3], Dhanashree Warad[4], Harshad Zawar[5] 1(Asst. Professor, Dept. of Information Technology, MMCOE, Pune, Maharashtra) 2, 3, 4, 5(Student, Dept. of Information Technology, MMCOE, Pune, Maharashtra)

Abstract- *In information systems, tagging is a popular way to categorize information and to search content. Therefore, almost all online newspapers, blogs, question-answer communities, and other similar sites make use of tags to categorize articles, posts, questions, answers, and so on . Automatic tagging processes content without human intervention, resulting in more standard and consistent results at lower costs.*

This paper gives a survey of different techniques used by the researchers for Question Tagging and Student Performance Prediction. An abstract view of the proposed system that we are going to implement helps to predict the expertise of an individual in the particular domain based on the attempted question from the question paper.

Keywords: ***Question-answering, Tagging, Neutral language processing, Parsing, Tokenizing, Stemming, Filtering*****.**

I. INTRODUCTION

Assessment is an essential part in determining the level of attainment of education. The expanded use of computer innovations has driven educational institutions and recruiters to search for approaches to utilize innovation in testing the student's ability [1]. Earlier, paper and pencil method is used as a standardized assessment method, but in recent days several Computer Based Testing (CBT) and Computer Adaptive Testing (CAT) methods are widely used[12,14]. In traditional testing, often students spend most of their time with questions that are not matching their knowledge and ability.

The CAT model is adaptively adjusting the approach and provides recommendation according to student's individual level is crucial to improving the learning efficiency. The term Tagging is a simple and efficient method to organize resources[11,13]. There are three types of tagging methods: (1) manual tagging, (2) semi-automatic tagging, and (3) automatic tagging [2, 4]. Manual tagging is the most commonly used method for organizing questions in the industry.

There are three main **advantages** to having an automatic tagging system.

- It can be used to tag existing questions that have less than five tags.
- Second, it can help new questioners by suggesting appropriate tags.
- Finally, it can warn questioners if they tend to select an inappropriate tag.

OVERVIEW

In this survey paper, we present a systematic detailed literature study of Automatic Question tagging and recommendation system. We initially list different user motivations and different kind of tagging web objects in this Section.

TAGS: WHY AND WHAT?

This section provides a detailed classification of user tagging motivations and also list different kinds of tags [4, 6, and 7].

Different User Tagging Motivations

- **Future Retrieval:** Users can tag objects aiming at ease of future retrieval of the objects by themselves or by others. Tags may also be used to incite an activity or act as reminders to oneself or others (e.g., the "to read" tag).
- **Contribution and Sharing:** Tags can be used to describe the resource and also to add the resource to conceptual clusters or refined categories for the value of either known or unknown audience.
- **Attract Attention:** Popular tags can be exploited to get people to look at one's own resources.

- **Opinion Expression:** Tags can convey value judgements that users wish to share with others (e.g., the "elitist" tag in Yahoo!'s Podcast system is utilized by some users to convey an opinion).
- **Task Organization:** Tags can also be used for task organization. E.g., "to read", "job search", "gtd" (got to do), "to do".

KINDS OF TAGS

- **Content-Based Tags:** They can be used to identify the actual content of the resource. E.g., Autos, Honda Odyssey, batman, open source, Lucerne.
- **Context-Based Tags:** Context-based tags provide the context of an object in which the object was created or saved.

  e.g., tags describing locations and time such as San Francisco, Golden Gate Bridge, and 2005-10-19.
- **Ownership Tags:** Such tags identify who owns the resource.
- **Subjective Tags:** Tags that express user's opinion and emotion, e.g., funny or cool. They can be used to help evaluate an object recommendation (item qualities). They are basically put with a motivation of self-expression.

  Organizational Tags: Tags that identify personal stuff. This type of tags is usually not useful for global tag aggregation with other users' tags.

  e.g., my paper or my work, and tags that serves as a reminder of certain tasks such as to-read or to-review.

II. LITERATURE REVIEW:

Automatic Question Tagging and Student Performance Prediction have gained a lot of interest from the user. Recent study on Question tagging and related work involves use of algorithms like RNN (Recurrent Neural Network)[5], SVM Label Classification, Random Forest Classifier (RFC)[5, 8], K-NN (K-Nearest Neighbor), Shadow Test Approach [5, 9]etc.

Gerel Tumenbayar and Hung Yu Kao help us in predicting the domain of the particular question posted. However this paper can be applied only when we have the tags for each and every question. Here they used Kruskal's algorithm to predict the domain of the respective question from the tags already present in the question. Here they have the tags already present in the test as well as the training dataset. With the help of Kruskal's algorithm they have created a Bayesian network from the training dataset and use it to predict the domain of the test dataset.

RNN (Recurrent Neural Network)

Fanming Dong, et al; have predicted the tags of a test datasets by training a model created using the training dataset that has got various question from different domains all together. Here tags were only provided in the training dataset and they have predicted the tags for the test dataset. After all the initial pre-processing they have first tokenized the dataset. After the tokenization they have converted each and every word in to vector values using Glove. After converting the words to vector they have formed their model using these words. After the formation of a model they have performed RNN (Recurrent Neural Network) [1,2,5] on it. Using this they have predicted the tags of the respective questions.

Advantages:

1) Recurrent neural network are even used with convolutional layers to extend the effective pixel Neighborhood.

Disadvantages:

1) RNNs are not able to keep track of long-term dependencies
2) It cannot process very long sequences if it uses tanh as its activation function
3) Gradient vanishing and exploding problems.

SVM (Support Vector Mechanism)

Avigit K. Saha, et al; have also predicted tags for the Stack Overflow Questions and have also tagged this particular question to their respective domain using the algorithm of SVM (Support Vector Mechanism)[1,10]. Here they have used a discriminative model approach. They have suggested that predicting the tags for a particular question dynamically can be a very difficult task and have suggested that not all the question may have the same way of predicting the tags as compared to some. So to provide a solution to the problem they have suggested a discriminative model approach with various kinds of solution of various types of domain oriented questions.

Accuracy: - they have got an accuracy of around 68.47%.

Advantages:

1) SVM is relatively memory efficient
2) SVM is more effective in high dimensional spaces.

Disadvantages:

1) SVM does not perform very well, when the data set has more noise i.e. target classes are overlapping.
2) In cases where number of features for each data point exceeds the number of training data sample, the SVM will underperform.

Sr. No.	AUTHOR, TITLE AND JOURNAL NAME	ADVANTAGES	REFER POINTS
1)	Harsh Parikh1 , Parth Patel2, "Question Tagging System",International Research	It can be used to tag existing questions that have lesser number of tags.	Word2Vec model
	Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 04 | Apr-2018 ",		
2)	Jyoti S Deshmukh , Amiya Kumar Tripathy "Text Classification using Semi-supervised Approach for Multi Domain "2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017)	The advantages in this approach: first, a joint classifier can produce pseudo labels for unlabeled data with high accuracy, which help train label-enhanced embeddings on a large unlabeled corpus	Produce Pseudo labels for unlabeled data
3)	R H Goudar' ,Shivanagowda+, Sreenivasa Rao," Design of Adaptive Question Bank Development and Management System" 2012 2nd IEEE International Conference on Parallel, Distributed and Grid Computing .	Adaptive question bank development and management system is intelligent system that takes care of fetching a balanced and standard question paper by distributing the questions equally among the levels.	Fetching a balanced and standard question paper
4)	, Jay Palan2, Ishita Shah3 " Automatic Question Generation from Paragraph "International Journal of Advance Engineering and Research Development Volume 3, Issue 12, December - 2016 .	Generate question based on subject verb object and prepositions present in the sentence by mapping it to certain predefined rules	Mapping of questions with certain rules
5)	Zhe Liu and Bernard J. Jansen ,"Understanding and Predicting Question Subjectivity in Social	1. By applying the classifier on a larger dataset, we then	Classifier to analysis the objective of

	Question and Answering",IEEE TRANSACTIONS on computational social systems, vol. 3, no. 1, march 2016	present in-depth analyses to compare subjective and objective questions, in terms of the way they are being asked and answered. 2. We find that the two types of questions exhibited very different characteristics, and further validate the expected benefits of differentiating questions according to their subjectivity orientations	question and answer by differentiating Questions according to their subjectivity orientations

III. PROPOSED SYSTEM

Users can Upload The Solved Question Paper to the system. The system preprocesses the input to Stop words and special symbol like (,?,/, ;, :)to get the fine data and extract the tag word from the data. The paper should contain questions from a different domain.

1. **Apply Preprocessing:**
   The system cans Understand Each Word from all the questions using Natural Language Processing (NLP). Basically, the natural language processing is used for extracting the features for understanding the sentence and word. They can analyze the words using two different ways like,

- **Sentence Tokenization-**

In this techniques system can divides the sentence into several tokens. It split the large raw text into several sentence to get more meaningful information out.
For eg.
"All work and no play make jack a dull boy, all work and no play".
The above sentence is divide into sentence like,
['All work and no play make jack dull boy.', 'All work and no play makes jack a dull boy.']

- **POS Tagging-**

i. This algorithm is used for detects if the word token is noun, verb, adjective
ii. POS Tagging in which a word is assigned in accordance with its syntactic functions. In English the main parts of speech are noun, pronoun, adjective, determiner, verb, preposition, adverb, conjunction, and interjection.

- **Word Tokenization-**

This technique the sentence or data can split into several words. For eg.
"All work and no play makes jack a dull boy, all work and no play". This sentence split into word like,
['All', 'work', 'and', 'no', 'play', 'makes', 'jack', 'dull', 'boy', ',', 'all', 'work', 'and', 'no', 'play'].

- **Word Lemmatization-**

Lemmatization is a more methodical way of converting all the grammatical/inflected forms of the root of the word. Lemmatization uses context and part of speech to determine the inflected form of the word and applies different normalization rules for each part of speech to get the root word (***lemma***).

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

- **Word Similarity-**
  By using this technique the system can find the similar words. We use the WordNet dictionary for finding the synonyms.
- **WordNet Dictionary-**
  WordNet is a combination of dictionary and thesaurus. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members.
  For eg.
  "last night"→ "yesterday"

- **Sentence Similarity-**
  By using this technique the system can find the similar sentence.

2. Text classification using Machine learning techniques

Naïve Bayes classifier is used for text classification. Firstly we train our model containing class and specific data of that class i.e. text document of the class and then we pass the context words as an input to our trained model to get center words which would be our tags. By applying the Naïve Bayes classifier we get the domain of the questions solved by the user from the question paper.

3. Analysis

Finally, the system will analyze the different users based on solved questions and give expertise for a particular domain. The system will also recommend the domain in which the user is lacking.

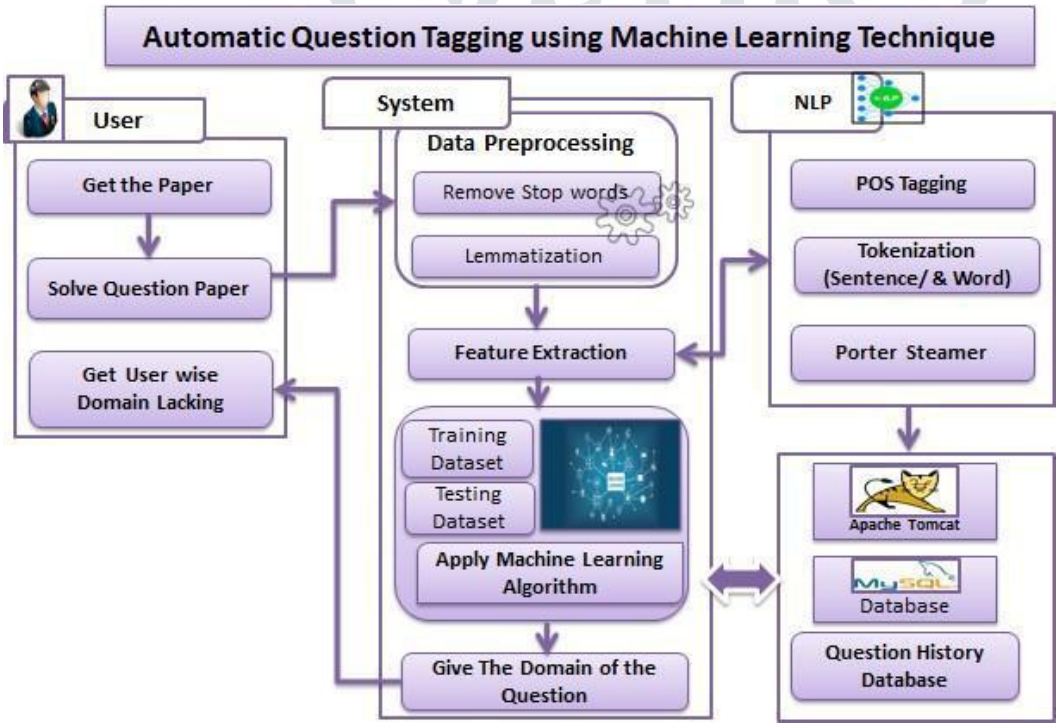

Figure: - Proposed System Architecture.

VI. ALGORITHMS USED

1. NLP (Natural language Processing)

We want a computer to communicate with users in their terms; we would not force users to learn a new language. This is particularly important for casual users and those users, such as managers and children, who have neither the time nor the inclination to learn new interaction skills. Many of the problems of AI arise in a very clear and explicit form in natural language processing and, thus, it is a good domain in which to experiment with general theories. So we are using NLP in our project.

a) **Porter stemming algorithm** (or '**Porter stemmer**') :
It is process for removing the commoner morphological and in flexional endings from words in English. Following are the steps of this algorithm:-

- Gets rid of plurals and -ed or -ing suffixes
- Turns terminal y to i when there is another vowel in the stem
- Maps double suffixes to single ones: -ization, -ational, etc.
- Deals with suffixes, -full, -ness etc.
- Takes off -ant, -ence, etc.
- Removes a final –e

2. Naïve Bays

Naïve Bayes Classifier is amongst the most popular learning method grouped by similarities that works on the popular Bayes Theorem of Probability- to build machine learning models particularly for disease prediction and document classification. It is a simple classification of words based on Bayes Probability Theorem for subjective analysis of content.

Use of Naïve Bayes Classifier:

- If you have a moderate or large training data set.
- If the instances have several attributes.
- Given the classification parameter, attributes which describe the instances should be conditionally independent.

In Naive Bayes classifier (NB), it is assumed that a term's occurrence is independent of the other terms. We want to find a class that gives the highest conditional probability given a document d:

d = Document

c= Class

$$\arg\max_{c \in C} P(c|d)$$

By Bayes rule [3],

$$P(c|d) = \frac{P(d|c) \cdot P(c)}{P(d)}$$

It is clear that

$$P(c) = \frac{|c|}{\sum_{c' \in C} |c'|}$$

And P (d) can be ignored since it is common to all classes.

There are two ways to compute P (d|c) based on the representation: either binary or term frequency-based. We show how to compute P (d|c) for the latter.

Let N_{it} - be the number of occurrences word w_t in document di, and V -

Vocabulary size.

Then P (di|c) is the Multinomial Distribution :

$$P(d_i|c) = P(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{P(w_t|c)^{N_{it}}}{N_{it}!}$$

P (|di|)|di|! Is also common to all classes and thus can be dropped. Finally, the probability of word wt in class c can be estimated from the training data:

$$P(w_t|c) = \frac{1 + \sum_{i=1}^{|D|} N_{it} P(c|d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c|d_i)}$$

Where D is the training data set.

VII. RESULT AND ANALYSIS

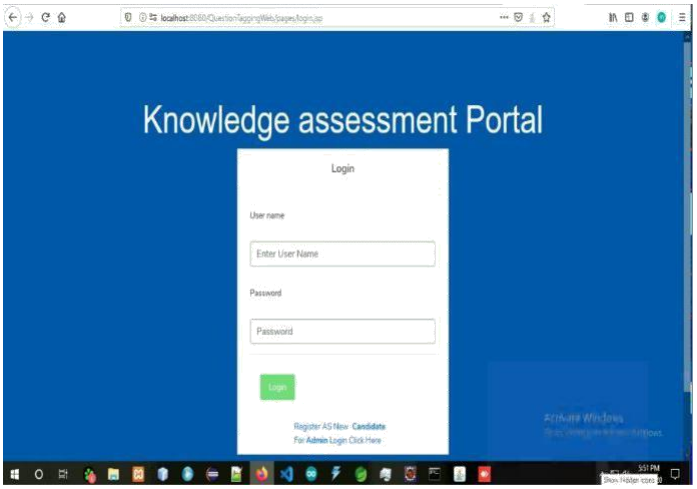

Figure 1

Figure 1-This page is the login page of our system. User can access their account by providing correct credentials. New Users can register to create their account.

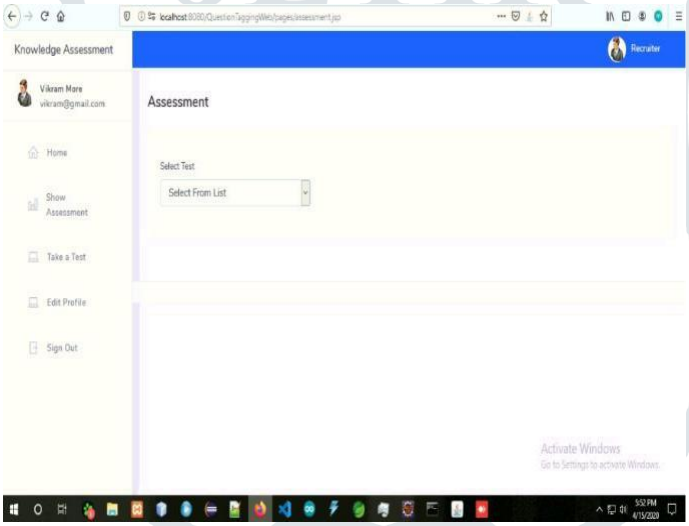

Figure 2

Figure 2- This page is the assessment page in which several options can be chosen .After login into account user can-

i) Appear for a test
ii) Edit Profile
iii) See Assessment
iv) Sign Out.

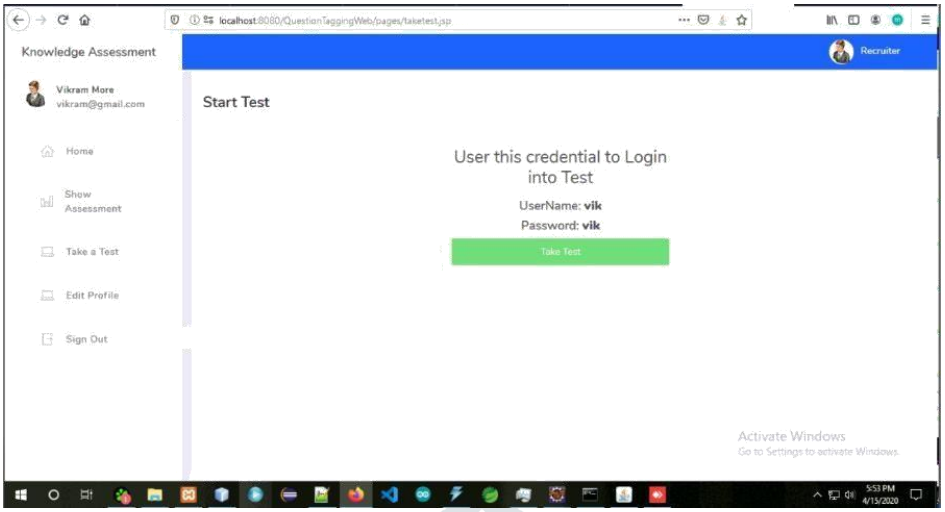

Figure 3

Figure 3-This page appears to start a test in an user's account. Credentials are provided to user.

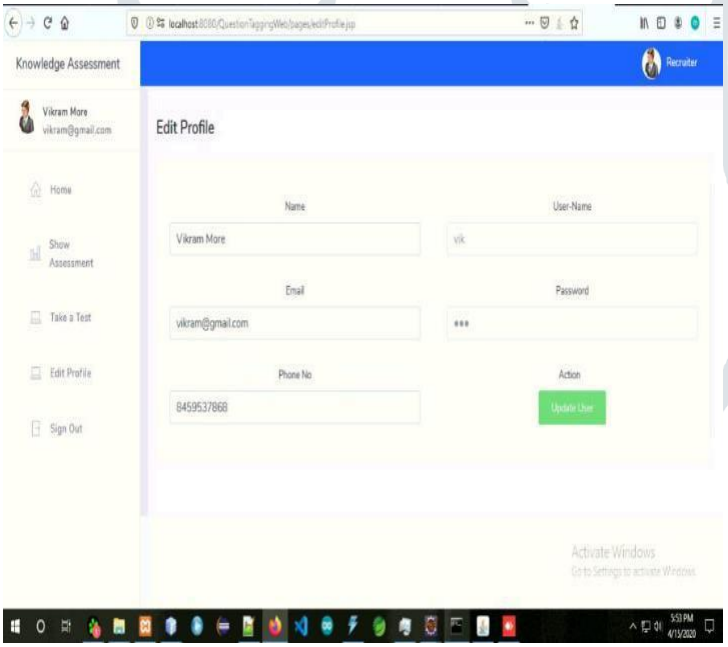

Figure 4

Figure 4-This page allows user to edit profile details of the account like Name, Email id, Phone no, Username, Password.

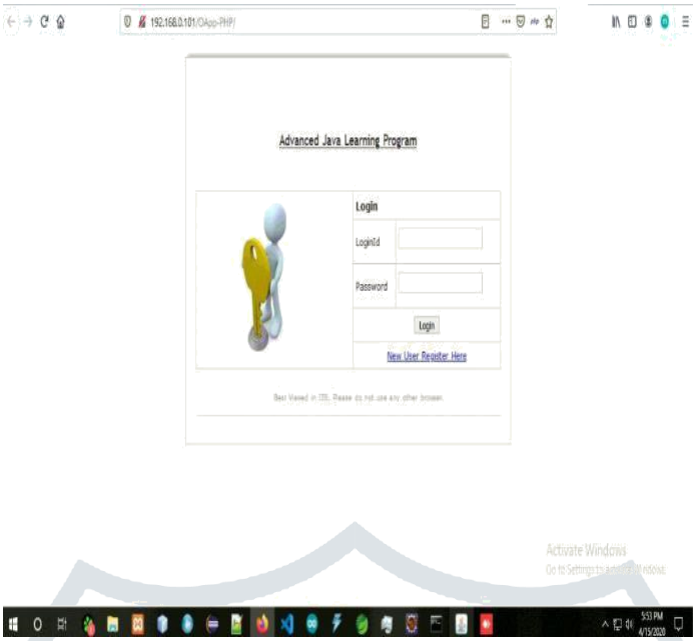

Figure 5

Figure 5-This is the login page to appear for a test of a particular subject. In this picture the subject is Advanced Java Learning Program.

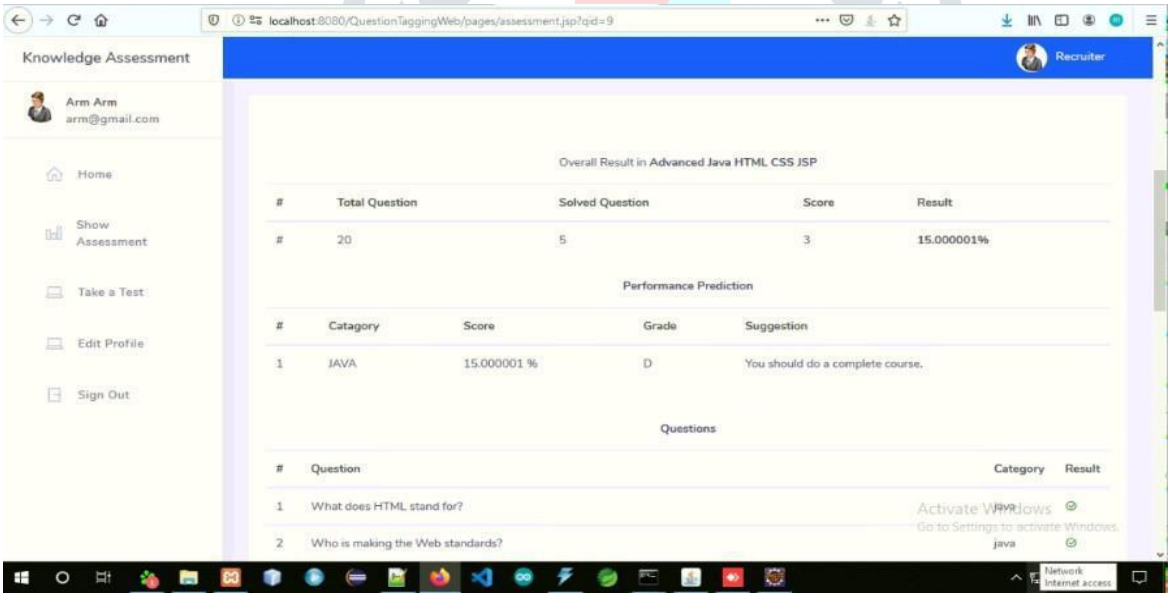

Figure 6

Figure 6-This is the result page in which the result of the test which the user appeared is shown. It shows
i) Total no. of questions in the test
ii) How many are solved by the user
iii) The score
iv) Percentage scored
v) Category of test
vi) Grade obtained by the user
vii) Suggestion about the performance.

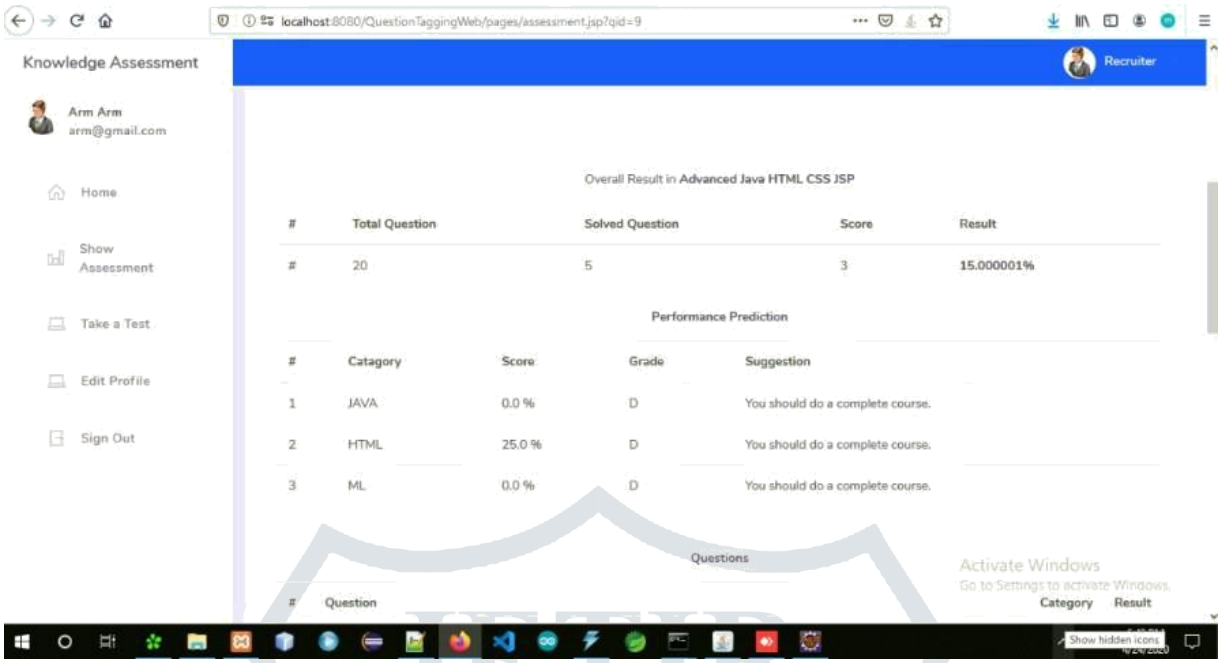

Figure 7

Figure 7- In this page overall result of the user is shown. All the result of the test which the user had appeared is shown here. In this page the result of Java, HTML,ML is shown. And suggestion are also given on the basis of the performance of the user**.**

VIII. CONCLUSION

This work analyzed the automatic Question Tagging and Student Performance Prediction System. This survey revealed that a significant amount of efforts have been made for predicting the domain of the question from the question paper. By doing this we will get the expertise for a particular user based on the solved question from the question paper.

Thus, observing overall literatures survey and respective contributions of different researchers, it can be visualized that evolutionary computing schemes can strengthen all the comprising functional components for question tagging and student performance prediction.

REFERENCES

[1] Bo Sun, Yunzong Zhu, Yongkang Xiao, Rong Xiao and Yungang Wei ,“ Automatic Question Tagging with Deep Neural Networks”, IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES 2018.

[2] Harsh Parikh1, Parth Patel2, Vatsal Sanghrajka3, Chintan Savla4, Manya Gidwani ,“ Question Tagging System”, International Research Journal of Engineering and Technology(IRJET)e-ISSN: 2395-0056Volume: 05 Issue: 04 | Apr-2018.

[3] Avigit K. Saha ,Ripon K. Saha†Kevin A. Schneider “ A Discriminative Model Approach for SuggestingTags Automatically for Stack Overflow Questions”, 2013 IEEE.

[4] Kamel Alreshedy, Dhanush Dharmaretnam, Daniel M. German, Venkatesh Srinivasan and T. Aaron Gulliver ,“ Predicting the Programming Language of Questions and Snippets of Stack Overflow Using Natural Language Processing”.

[5] Tumenbayar, Gerel, and Hung Yu Kao. “Topic Suggestion by Bayesian Network Enhanced Tag Inference in Community Question Answering.” 2016 Conference on Technologies and Applications of Artificial Intelligence (TAAI), 2016, doi:10.1109/taai.2016.7880110.

[6] Dong, Mao and Zhu. “Transfer Learning on Stack Exchange Tags.” Stanford University.

[7] Deshmukh, Jyoti S, and Amiya Kumar Tripathy. “Text Classification Using Semi-Supervised Approach for Multi Domain.” 2017 International Conference on Nascent Technologies in Engineering (ICNTE), 2017, doi:10.1109/icnte.2017.7947982.

[8] Charte, Francisco, et al. “QUINTA: A Question Tagging Assistant to Improve the Answering Ratio in Electronic Forums.” IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), 2015, doi:10.1109/eurocon.2015.7313677.

[9] Saha, Avigit K., et al. “A Discriminative Model Approach for Suggesting Tags Automatically for Stack Overflow Questions.” 2013 10th Working Conference on Mining Software Repositories (MSR), 2013, doi:10.1109/msr.2013.6624009.

[10] Hua-Hua Chang and Zhiliang Ying, “a-Stratified Multistage Computerized Adaptive Testing With b Blocking”, in Sage Publication, Educational Testing Service, Vol. 25 No. 4, December 2001, 333–341

[11] Wim J. van der Linden, “Constrained Adaptive Testing with Shadow Tests”, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8 2.

[12] Shital Kakad, Prachi Sarode, JW Bakal,"A survey on query response time optimization approaches for reliable data communication in wireless sensor network" International Journal of Wireless Communications and Networking Technologies, 2012.

[13] Shital Kakad, Sarode, J Bakal," Analysis and Implementation of Top k Query Response Time Optimization Approach for Reliable Data Communication in Wireless Sensor Networks" International Journal of Engineering and Innovative Technology (IJEIT) 2013.