Career Advice Professionals Recommendation System

Aishwarya Sapakle
UMBC
Computer Science
asapkal1@umbc.edu

Arti Singh
UMBC
Computer Science
artis2@umbc.edu

Rohit Kumar
UMBC
Computer Science
rohitk1@umbc.edu

Snehika Pandey
UMBC
Computer Science
spandey3@umbc.edu

Abstract

In current world there are many online question answering communities. Career Village.org is one among them, which provides career guidance to 3.5M online learners in every field, volunteered by 25000 professionals, till date from respective fields. This online platform receives large volume of questions from users, and it becomes difficult to direct them to most relevant group of professionals who will most likely answer those questions. So, in this project we plan to implement a Question Mapping system to map the question to the most pertinent corps. We plan to build this system by first clustering all the volunteers based on profile details of the professionals. Then by using ranking algorithm we will increase the chances of question hitting the professionals who are most likely to answer. We will evaluate the correctness of the proposed system by using the mean reciprocal rank. .

1 Introduction

Career Village.org, founded in 2011 at NY City, is an online question answer platform which crowd sources career advice to students who cannot reach-out to counselors in-person. This is akin to known platforms like Stack Overflow, Quora or Chegg expert QA, which provides guidance by answering the queries of each individual.

This system currently has 25000 professionals profiles. These professional have opted to answer question from their field of expertise. Each professional is asked to provide a title to their profile along with their current industry and tag. Tags are keywords in the system which helps to find the relevant question to be directed to the particular person. Tags are also entered when a user posts a question on the system, so as to compare it against the expert profiles and notify accordingly to the volunteer about the post.

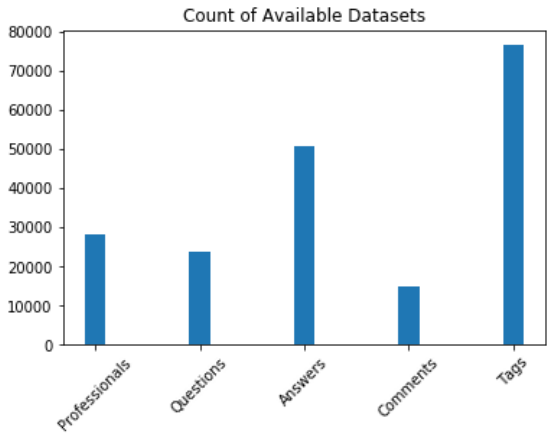

Figure 1: Plot of Available Data Count for Each Dataset

Currently the questions are posted in open for anyone from that field to answer it. The main concern in this system is to direct a question such that there is a high probability of the question getting answered. The central problem of any QA system is question routing to an expert and getting the best answer.

2 Proposed Solution

As there is a large pool of professionals, the first hurdle is to deal with the issue of segregating them into related fields with their industry. This is because the probability of answering a question related to the experts industry is more considering the professional experience and knowledge with that field. In order to sole this, we use similar questions approach i.e. we find out which professionals have answered similar kind of questions before and then predict based on ranking who is most likely to answer the question.

For each industry, we apply a ranking algorithm based on the professionals history records to find the experts, who is most likely to answer. This

will help us create a prediction model based on the ranking of the professionals to route the questions accordingly.

3 Related Work

There are many similar QA platforms like Quora, Stack Overflow, Yahoo! Answers, etc. We have discussed underlying algorithms for some of them below.

1. Quora an online platform to gain and share knowledge. As per a blog on Forbes Quora uses Machine Learning algorithms for question understanding and further classifies it based on types and labels it to determine its respective topic. If a user is looking to post a question, Quora displays relevant answered questions by a ranking algorithm. Quora has an A2A (Ask to Answer) feature which finds the list of relevant experts, from which the users can select to whom to direct question. Quora uses algorithms such as logistic regression, elastic nets, random forest, deep neural networks, vector models and other NLP techniques. They also use K-Means and other clustering approaches along with their own QMF (Quora Matrix Factorization).

2. Stack Overflow a widely known platform for topics mostly related to computer programming languages. Currently categorizes a wide range of topics using appropriate tags wherein the user has to enter the tags manually based on their understanding. The only drawback is it has a large database of tags which often makes the whole process to categorize the question correctly to a tag. To overcome this, they have come up with a hybrid auto-tagging feature which detects the tags for the programming language used in the question posed by the user. For this, they use the SVM (Support Vector Machine) classification. As stated for Stack Overflow, they use a huge database of tags to compare the questions, we are also using a database of tags but our database is constantly updating with new tags as new questions are being answered.

3.Yahoo Answer - Yahoo being one of the largest knowledge exchange community with largest database, provides a multipurpose platform which is being used for technical knowledge exchange as well as for non-professional talks. This is achieved based on classification of categories using K-Means Clustering and yielding result on the basis of intuition.

4 System Architecture

The problem statement says that when a new question comes in, it has to be directed to the professionals who are most likely to answer it. The question entered by the user will have associated tags specified by users itself while posting. Here, we plan to take these tags into consideration to build our system architecture based on the Content-based recommendation approach6. In this system, we retrieve the set of tags for each particular industry which are used to map incoming questions to the most appropriate professional by comparing it with the user specified question tags. This approach will map the question to the professional with a higher probability and accuracy of answering it. In order to do this, we first create groups of tags associated with different industries. This mapping of tags to industries is shown in the Figure-2.

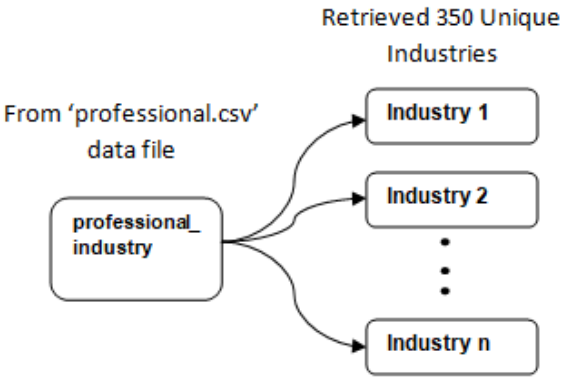

Figure 2: Retrieve 'n' Unique Industries.

First, we mapped professionals into respective industries based on the experience and knowledge of professionals that they have in the respective industry. Then, we find out the questions already answered by each of those professionals to train the system with the set of questions for each professional under each industry. We then make a group of tags associated with each of these questions to form an industry-to-tags mapping i.e. we get a set of tags for each of the industries.

Based on this industry-to-tag mapping we then match the tags of the new questions to the tags that we have found earlier to predict the corresponding industries. Then a ranking is performed on the professionals under the selected industries using a ranking algorithm and the professionals with good ranking scores are sent the email for answer-

ing the new posted question. In order to test the robustness of our built-in system we experimented it with two other set-up based on various factors and configuration named as Experiment 2 and Experiment 3.

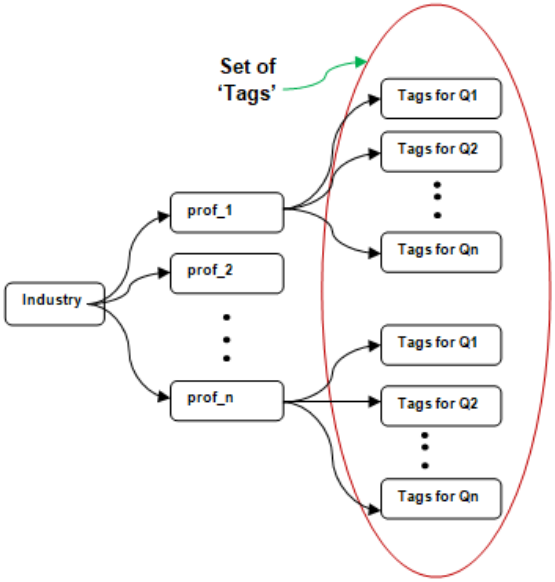

Figure 3: Industry Mapping to Set of Tags.

5 Method

Below are the some of the methods and experiments we used to test this model.

5.1 Experiment 1

As there is a large pool of professionals, the first issue to deal with is segregating them into related fields as the probability of answering the questions related to the experts industry is more, as that professional would have more knowledge about that field. To go about this we are using similar questions approach i.e. we find out which professionals have answered similar kind of questions before and then Predict who is more likely to answer the question now.

For each industry, we will apply a ranking algorithm based on the professionals history records to find the experts, who are most likely to answer. This will help us create a prediction model based on the ranking of the professionals to route the questions accordingly.

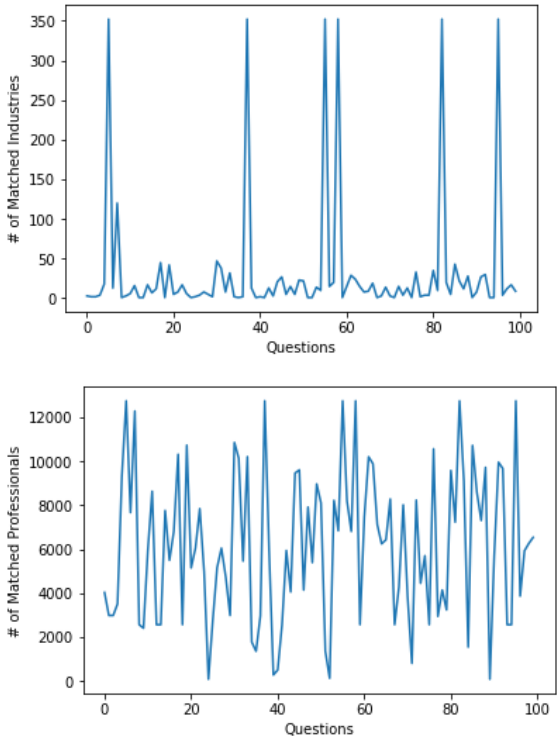

Figure 4: Experiment 1 - Plot for Number of Matched Industries and Professionals for Each Question (along x-axis)

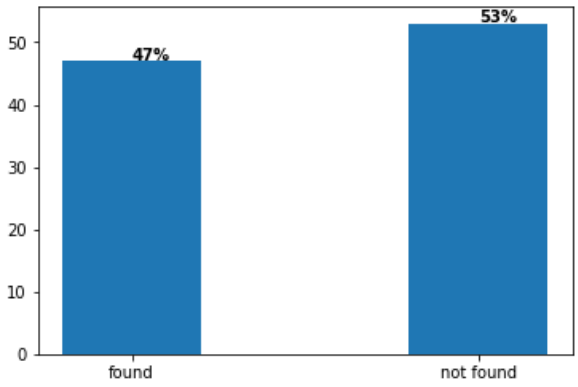

Figure 5: Experiment 1 - Evaluation (True Positive and False Negative)

5.2 Experiment 2

In this set-up, we are not only just taking the user specified tags but also extracting our own defined tags using the questions body and title then using it to map to the appropriate professional as done above. This is because the tag entered by user might not be correct. All other system configuration and set up, we kept same as the earlier set-up. This extraction of tags from the questions body and title is what we can call as the meta-data of data i.e., meta-data of question to test the built

system as a meta-data can provide more detailed and accurate result. For the implementation, we have used some of the built-in sklearn extraction features such as Count Vectorizer, stop-words etc.

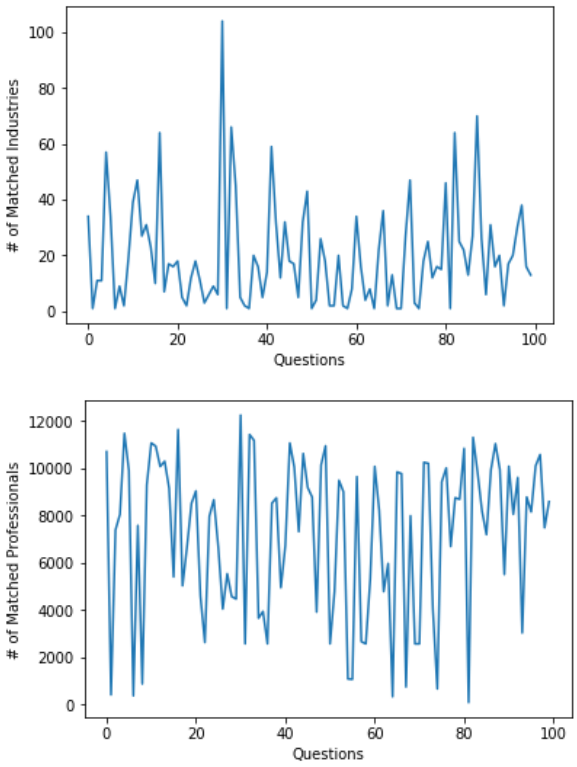

Figure 6: Experiment 2 - Plot for Number of Matched Industries and Professionals for Each Question (along x-axis)

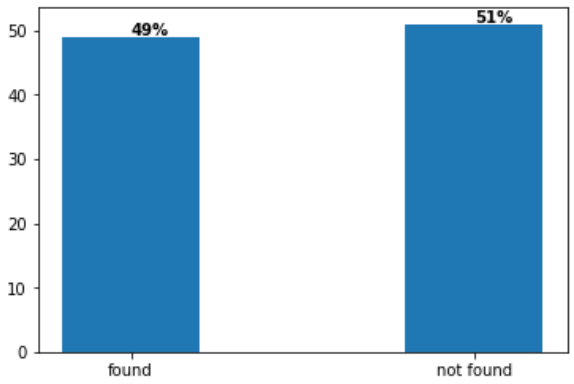

Figure 7: Experiment 2 Evaluation (True Positive and False Negative)

5.3 Experiment 3

Here, we have taken the answering frequency of professionals in consideration to find how likely and quickly a professional can answer the incoming questions and based on that we map the questions to the appropriate professional. Basically, this answering frequency is used for ranking which lists the professional in increasing order of their likeliness to answer any question. Ranking based on answering frequency is calculated as the ratio of the number of questions answered to the number of months a professional is registered with the organization.

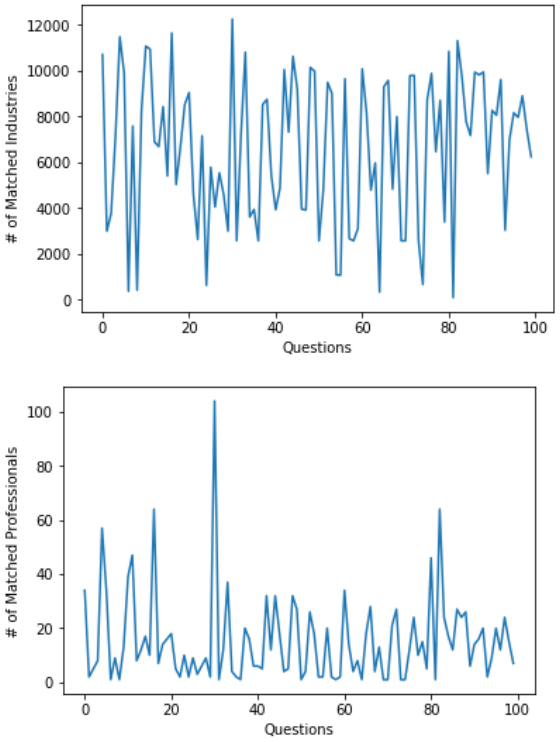

Figure 8: Experiment 3 - Plot for Number of Matched Industries and Professionals for Each Question (along x-axis)

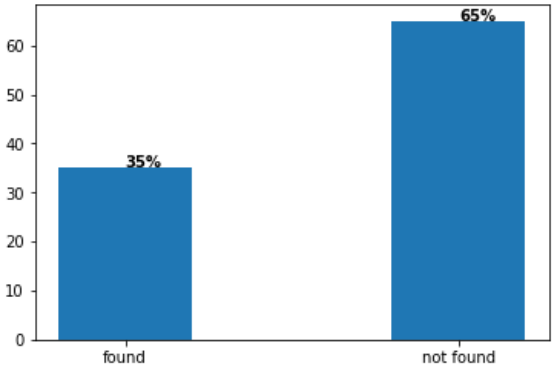

Figure 9: Experiment 3 Evaluation (True Positive and False Negative)

5.4 Experiment 4

In this experiment, we are considering the semantics of the words extracted which is different from the above experiments performed in the way that earlier we have performed matching based on the tags. For the implementation, we have extracted keywords using Tf-Idf Vectorizer based on the questions titles from training data. Then, we created a sparse matrix of the keywords and questions titles. Further, we reduced the dimensionality of the sparse matrix obtained above using Truncated SVD (LSA) to form 2-D eigen vectors for each of the questions in the training data. For correct predictions of each question in the test set, we extracted the keywords/tokens from the question title and corresponding sparse matrix based on the token. We formed a 2-D Eigen vector of this sparse matrix followed by calculating the cosine distance between the above calculated Eigen vector and the Eigen vectors calculated for each question from given question dataset. We took the training questions having minimum distance from the test question Eigen vector which infer that semantically these questions from the training data are similar to the test question, so the professionals who answered these questions have a higher probability of answering the test question. Based on our visualization of the output, the similarities between the questions were real accurate sematically. However, this approach is not fully completed, we plan to implement this as a future work to further cluster the questions and industries based on semantic analysis here. It is our belief that this approach will lead to a more robust and accurate solution. However, it needs more analysis and design discussion.

6 Evaluation

Evaluation of the experiment was done by computing the accuracy of the model i.e., we have computed the true-positive and false-negative, which describes the status of the Actual Outcome and the Predicted Outcome.

In Experiment 1, the match percentage of the Actual outcome is observed to be 47 percent. This provides the true-positive equal to 47 percent and a false-negative of 53 percent.

Further, Experiment 2 provides the match percentage of the actual outcome as 49 percent, letting us to conclude that the true-positive is equal to 49 percent and a false-negative of 51 percent.

In Experiment 3, we have observed that the match percentage is 35 percent of the Actual Outcome which concludes the true-positive to be 65 percent and gives false-negative 53 percent. Lastly, in Experiment 4, we have analyzed the semantic similarities of the questions by converting and compressing the questions into Eigen vectors and computing the cosine distance between the questions. This provide us a different approach to map the questions based on its semantic meanings, left for future work.

7 Result

The above experiments and observations concludes that, when we used the metadata of provided questions table to direct a question to an appropriate professional, as in Experiment 2, the observed accuracy is higher as compared to the other methodologies used.

8 Conclusion

Career Village is a large and diverse question answer community medium for youngsters to communicate with the Industrial and Subject Expert, and a place to seek advice, gather opinions, and satisfy ones curiosity about things which may not have a single best answer. We have utilized the existing professional data to built the cluster of Industry and get the list of professional for each Industry. This helped us to map the Questions to most likely professional of the Industry. To get expected result, we have used Ranking Algorithm considering the professional's activity, which is derived using the Question metadata, Answer metadata and Professional metadata. Finally, we have conclude that using the metadata of Questions and Professional against the baseline experiment, we are able to achieve a better performance for our model. Furthermore, we have implemented and analyzed the Latent Semantic Analysis (LSA) on the Question metadata Using TF-IDF Vector and Cosine Similarity to conclude that we can have some future work using the semantic approach as the base.

Acknowledgments

We would like to thank and express our sincere gratitude to Dr. Frank Ferraro for constant support and motivation starting from the topic selection to the implementation part where he has provided valuable guidance on how to approach the

problem by suggesting methodologies and journal support.

References

[1] CareerVillage - *https://www.kaggle.com/c/data-science-for-goodcareervillage*

[2] Zhou Zhao ; Lijun Zhang ; Xiaofei He ; Wilfred Ng. 2015. *Expert Finding for Question Answering via Graph Regularized Matrix Completion. IEEE Volume 27(4): 993 1004*

[3] Kenneth Ward Churc,Patrick Hanks. 1990. *WordAssociation Norms, Mutual Information, and Lexicography*

[4] Forbes Article, *https://www.forbes.com/sites/quora/2017/04/19/how-does-quora-use-machine-learning-in-2017/64c03fff3f3a*

[5] V. Smrithi Rekha,N. Divya,P. Sivakumar Bagavathi. 2014. *A Hybrid Auto-tagging System for StackOverflow Forum Questions*

[6] Recommender Engine for CareerVillage. *https://www.kaggle.com/ididur/nn-based-recommender-engine*

[7] Peter F. Brown,Peter V. deSouza,Robert L. Mercer,Vincent J. Della Pietra,Jenifer C. Lai . 1992. *Class-Based n-gram Models of Natural Language*

[8] Lada A. Adamic1, Jun Zhang1, Eytan Bakshy1, Mark S. Ackerman. 2008. *Knowledge Sharing and Yahoo Answers: Everyone Knows Something*

[9] Joran Beel, Bela Gipp. 2015. *Google Scholars Ranking Algorithm: An Introductory Overview*

[10] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. *Item-Based Collaborative Filtering Recommendation Algorithms*

[11] Thomas Hofmann. 2017. *Probabilistic Latent Semantic Indexing*