Enhanced Search Replies Ranking In Crowd Sourced Data Application

Juee Gosavi[1], Prof.B.N.Jagdale[2]

Department Of Information Technology, MIT College Of Engineering, SPPU, Pune (India)

[1]jueeg1693@gmail.com

[2]bjagdale@gmail.com

***Abstract*— Nowadays, Social communities are growing rapidly where opinions are expressed in natural language independently. Current question answering forums are a representation of future web search services like searching and posting queries or answers. Finding the significant answer to a recently posted query is expected by a framework, it furnishes the pool of answers with similar questions links, which could be an extended task. A solution to this problem is the framework provides a way to effectively rank the most relevant and best answers which are from historical archives based on similar queries found. It constructs training samples with positive, negative and neutral classes and then online component retrieves similar queries with their answer pools. Two approaches were compared to retrieve similar questions. The objective is to rank answer candidates based on pairwise comparison where question-answer pairs are ranked using an SVM-based rank approach based on an offline trained model which provides the user with most relevant answers for a given posted question.**

***Keywords*— Answer Selection, Community Question Answering, Ranking, Question Classification, Data Mining Algorithms**

I. Introduction

The leading form of knowledge retrieval is represented by question answering system which is recognized by user information requirements that are relatively conveyed in terms of natural language sentences or queries and is a type of natural sorts of human-computer interaction. Apparently with conventional data recovery, in question answering where whole contents are viewed as relevant to the asked data, as a reply to a query precise parts of data are returned. Finding a brief and comprehensible with correct answer, which refers to passage, word, picture, sentence, an audio fragment, or whole document, is required by users of a question answering system .

Online social networks growing rapidly nowadays which offers a wide range of options to express opinions in natural language. Question Answering Systems are one of the web forums which allow users or experts to ask or to answer to a question in natural language. The main functionality of question answering system is, for a given question from web and collection of documents, find the exact answer which satisfies the user [8].Community Question Answering (CQA) sites have seen a spectacular increase in popularity in the recent years. With the advent and popularity of sites like Yahoo! Answers, Cross Validated, Stack Overflow, Quora, HealthTap more and more people now use these web forums to get answers to their questions. These forums offer individuals the flexibility to post their queries online and have multiple experts across the world to answer them, whereas having the ability to produce opinions or expertise to help other users, a quality of answers encourages more participation and recognition .

Multiple ways are present to find answers which are relevant to the question asked but for a particular query, long lists of probably relevant documents are returned by current community question answering sites without identifying the vital of the result with a brief answer. Therefore the most essential tasks for knowledge consumers or users is to identify precise answer data, that is getting direct specific relevant answer for a query. Proper organization of required knowledge is needed by returning exact and specific answers [10].

In Community question and answer systems when we try to find answers to the questions we use archives where we can find them using theoretical base. But it can be time-consuming part to find out questions and where they can be associated with different answers and to find out relevant answers they need to go through a lot of answers to find what is needed [1].It is necessary to find out a precise answer for a given query which most relevant and recent. Also, information seekers need to wait for a long time for receiving an answer from other users, so finding similar questions and answers from historical documents will help to reduce time. A social network that presents an option to conventional web findings is recognized as Community Question-Answering (CQA) forums. Users of forums enters their required information as a proper questions in natural language and get straight replies written by people or experts ,instead of retrieving results of web search networks. Natural language contents comes in various qualities like questions-answers scopes from supreme quality content to low grade content to unrelated content or even offensive content. Due to which complications are increased in voting given to best answers and selection of supreme quality answer becomes most important [6].

The Question Answering system contains three methods as Question Classification (QC) which is a machine learning classifier, used for identification of the type of answer related to the input query based on training. The type of answer helps to identify a corresponding context-ranking model for ranking retrieved documents, Document, and Passage Retrieval(DPR) uses a Boolean oriented method for document recover and the density oriented method to identify related documents. Restricting received documents size to three sentences is important to create the basic answer component and the Context Ranking Model [12].To retrieve knowledge online, Community Question Answering (CQA) forums are the suitable option, with the facts like a. Information requester can post their questions based on any domain and receive answers presented by other user's .b. By using community exercise, they are obtaining better answers than search engines. As compared to automated CQA systems, CQA usually obtains better quality answers as they are based on human intellect and perception. c. Endless sets of QA pairs are gathered in their archives, which encourages the perpetuation and searching answered queries [10].

II. Review Of Literature

In community question and answer systems, we have a tendency to realize answers we use archives where we can realize them using theoretical base. However, it can be time- consuming part to seek out queries and where they can be associated with different answers and to find out relevant answers they need to go through a lot of answers to find what is required. To overcome this problem following papers were referred:

Paper [2] shows the empirical study and analysis of the answers to predict acceptability of the answers by the asker or community user by using Bayes classifier model, the approach identifies topic modeling to extract features for pattern identification for selecting best answers. This approach analyzes StackOverflow Q&A for selecting best answer whereas method in [3] represents an automatic content migration from web forums to latest Question answering forums based on binary classifier built upon text features for identifying best answers with good performance. Results provide a positive approach against automatic migration of crowd sourced information from legacy forums to latest question answering sites.

The author in this study [4] proposed an approach based on TF-IDF which identifies the best answerer for a newly posted question which contains vector space model where user's interest and user's expertise are considered while selecting best answerer. Based on historical question archives using Latent Dirichlet Allocation (LDA) answerers interests are modeled. The framework used in the paper [5] is based on the expertise level of the answerer in CQA session using SVMbased and Ranking SVM-based methods, questions are ranked and routed to the appropriate answerer. In results, Ranking SVM-based methods perform better than SVMbased methods on real-world datasets. In another study [6], the system uses a perceptron and a ranking-SVM based method in the stochastic gradient descent (SGD) framework, for regularization and learning from noisy data for the improvement in the answer ranking for social QA with the use of feature engineering along with learning algorithms. Results show the usefulness of query expansion techniques as well as the effect of regularization at the time of learning from noisy data.

In paper[10] the author proposed a three-level scheme which is based on generating a query-oriented summary format answer in form of novelty and redundancy. It calculates the global ranking score and combines it with a relevance score. It is based on calculated global ranking scores, which uses two various methods to build top K no. of answer set, and then solves an optimization problem to generate as a summary of top answers to a question asked by the user.

III. Proposed System

The Main idea of this system is to effectively rank answers which are most relevant and best from historical archives based on similar queries found. The system architecture comprises of modules which are an offline module and an online module which utilizes K nearest neighbor algorithm compared with Naïve Bayes algorithm for similar question retrieval and SVM-based rank model for finding most relevant answers for the newly searched question in which real-time HealthTap.com dataset is utilized as a part of training and testing data. The dataset comprises of queries asked by various patients which are replied by various specialists and user ratings given by different users of HealthTap.com community. The precision and recall values are compared for retrieving similar questions along with the most relevant answers.

A. Offline Module :

This module randomly selects questions from the training set and establishes positive, negative and neutral training samples in the form of preference pairs which are based on user ratings and user views. In this component based on preference pairs, the rank model is trained. The pairs are in the form of positive, negative and neutral pairs in terms of ratings or voting given by users or asker of the question. Asker rates answers according to its relevance to question. These ratings are stored in datasets for future relevance findings.

B. Online Module :

In online component user searches or posts any question, first it pre-processes the question and then it searches similar question and answers from repositories.

1) *Pre-processing Module:* Here, the asked question is processed by removing stop words and keyword extraction. Natural language processing is applied to process question and tag question features using part of speech tagging.

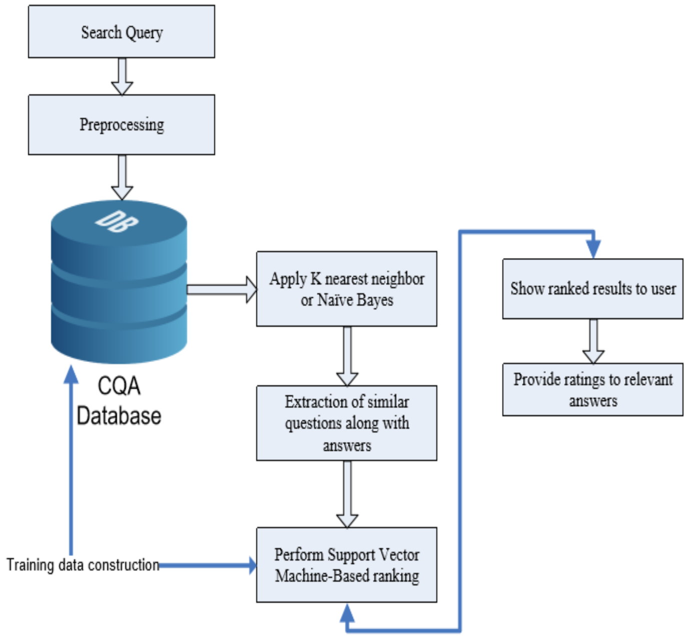
Search Query
Preprocessing
DB
CQA Database
Apply K nearest neighbor or Naïve Bayes
Extraction of similar questions along with answers
Perform Support Vector Machine-Based ranking
Training data construction
Show ranked results to user
Provide ratings to relevant answers

Figure 1. Proposed System Architecture

2) *Extracting relevant questions and their answers:* Based on the syntactic relevance of similar questions with their answers are retrieved from the database. For this process, KNN strategy is used to find top similar questions Also Naïve Bayes is performed to retrieve relevant questions to posted query or searched query and to compare results with KNN algorithm.

3) *Learning to Rank Model*: Learning to rank model sorts all the answers related to the returned similar questions. It is already trained in offline learning and it takes question answer pairs from online components which provide a ranking to pairs. It uses the Support Vector Machine based algorithm for finding best relevant answers [1].

C. Evaluation Method

For the experimental result, we have notations as follows:

TP: True Positive (number of instances which correctly retrieved),
FP: False positive (number of instances which incorrectly retrieved),
TN: True negative (correctly retrieved the number of instances as not required)
FN: False negative (incorrectly retrieved the number of instances as not required),

On the basis of this parameter, we can calculate two measurements

1. Precision : TP + FP != 0
   Precision = TP / (TP + FP + 0.1)
2. Recall : FN = FP - TP;
   TP + FN != 0
   Recall = TP / (TP + FN + 0.1)

IV. Algorithm

For the similar questions retrieval, we have analysed two algorithms K nearest neighbor and Naïve Bayes algorithm. After that output questions and answers are ranked using a support vector machine method.

A. KNN Algorithm

The K-Nearest Neighbor (KNN) is a machine learning algorithm whose goal is to classify objects into predefined classes of a sample group of documents. There is no need of training data to implement classification in the KNN algorithm, this data can be used during the testing task. This algorithm is based on identifying the most relevant questions from sample groups of documents. For the retrieval of similar questions associated with the user query first question is processed and features are extracted. Using TF-IDF method distance calculated with the query features. Term frequency-inverse document frequency is a measurement method which allows the calculation of the score for each word in every document. The technique finds the weight which evaluates the significance of terms in a collection of documents. The significance of the content is increased relative to the number of occurring contents in the documents.

B. Naïve Bayes Algorithm

Naïve Bayes classifier is based on the probabilistic model and depends on the Bayes theorem. In the supervised learning, the Naïve Bayes classifier work. The particular attributes are considered.
Pr(a/b) = Pr(a/b)* Pr(a)/ Pr(b)
Pr(a/b) = rear probability
Pr(a) = preceding probability of class.
Pr(b/a) = potential probability of class
Pr(b) = preceding probability of predictor.

C. Support Vector Machine – Based Rank Model

In machine learning, support vector machine represents supervised learning methods with related learning algorithms. It determines data used for classification and regression analysis. Support Vector Machine helps to retrieve best and most relevant questions from the input set of similar questions. Based on voting answers for the particular questions are ranked and most relevant questions with their answers are selected.

V. RESULT AND ANALYSIS

We have collected more than 35,000 questions and answers from websites to create Question Answer dataset which contains multiple answers with respect to questions along with their ratings. The table shows resulting values for user query "What are symptoms of blood cancer?" which represents K nearest neighbor returns less no. of relevant questions this gives more specific and precise results compared to naïve bayes algorithm also time and memory required to process questions are more with this algorithm.

TABLE I
COMPARISON CHART FOR KNN AND NB

Keywords	Algorithms	Similar Question Count	Best Question Count
Symptoms, Blood, Cancer	K nearest neighbor	5287	12
	Naïve Bayes	6961	29

Table 1.Comparison of two algorithms in terms of no. of relevant questions returned according to the user's example query.

The performance is evaluated in terms of precision value and recall value where recall is a performance measure of the entire positive section of a dataset and precision is a performance measure of positive predictions. In results, it shows that K nearest neighbor along with SVM performs better compared to the Naïve Bayes algorithm.

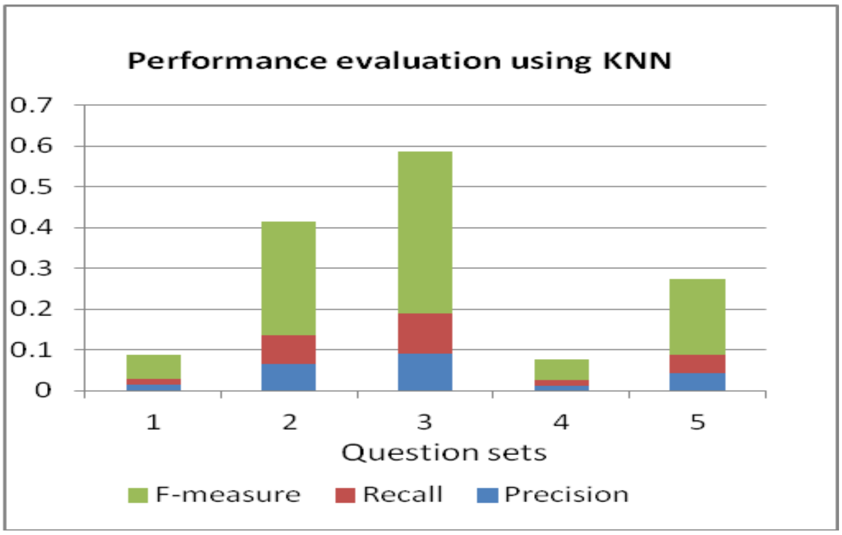

Figure 2. Performance evaluation in terms of precision and recall using KNN algorithm

Figure 2. shows precision and recall values for three different questions when applied K nearest neighbor algorithm for retrieving similar questions and SVM for retrieving best question along with answers whereas Figure 3 shows precision and recall values for three different questions when applied the Naïve Bayes algorithm for retrieving similar questions and SVM for retrieving best question along with answers. The graph shows K nearest neighbor gives better performance compared to the Naïve Bayes algorithm.

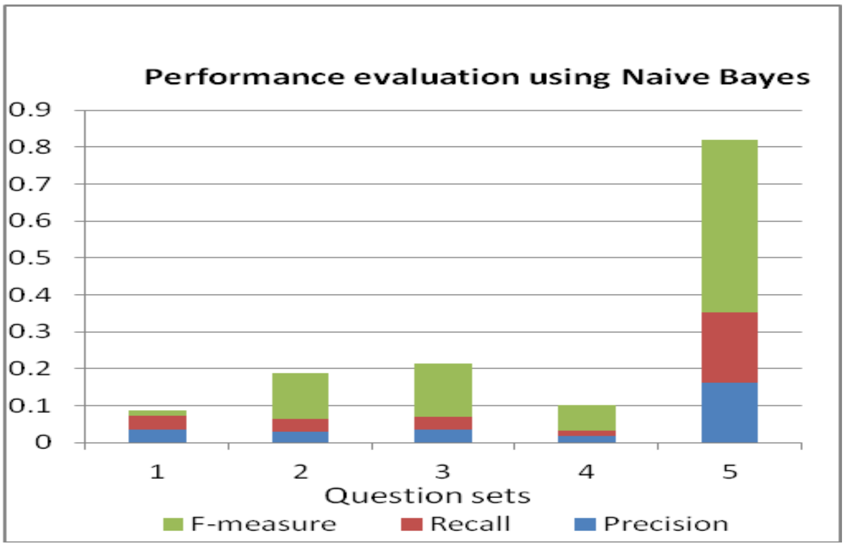

Figure 3. Performance evaluation in terms of precision and recall using Naïve Bayes algorithm.

The time graph represents the time required to retrieve a number of questions using different algorithms with respect to system count. According to graphs as shown in figure time required to process and retrieve questions along with answers using K nearest neighbor is less compared to naïve bayes algorithm. Time is calculated in terms of Millis against the system running count.

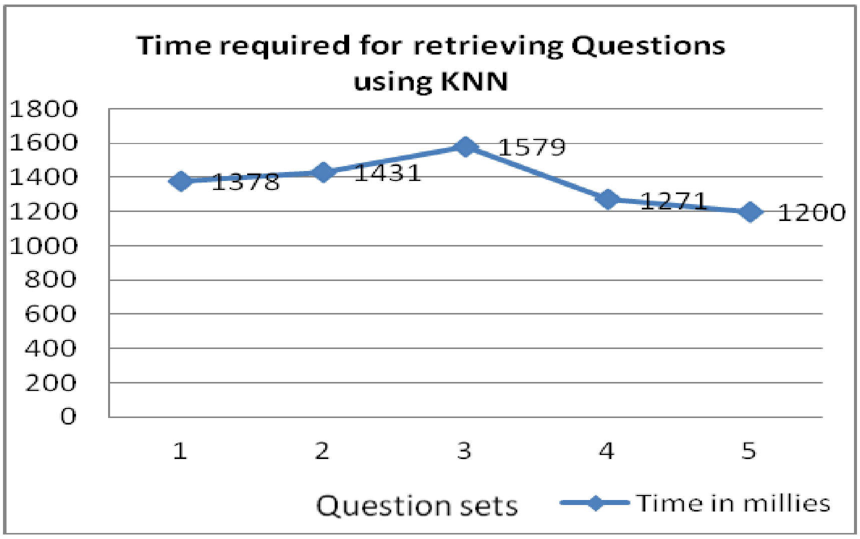

Figure 4. The time required for retrieving Questions using K nearest neighbor algorithm.

Also, no. of similar questions retrieved with knn are less compared to naïve bayes which proportionally requires less time and memory.

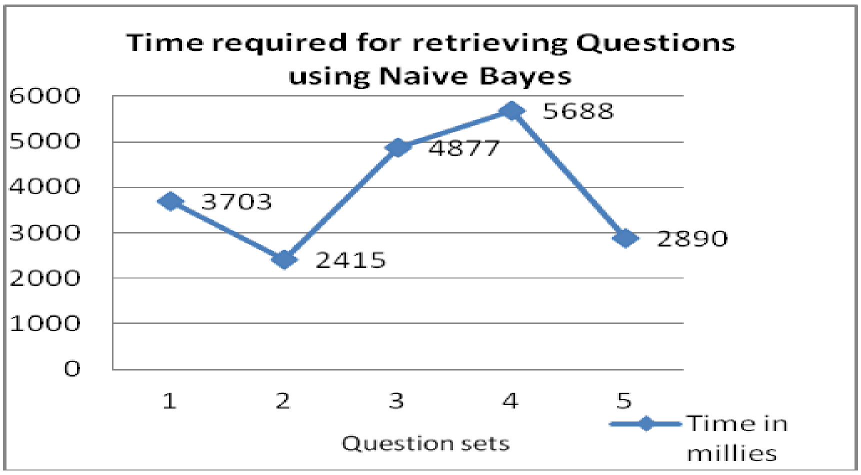

Figure 5. The time required for retrieving Questions using the Naïve Bayes algorithm.

VI. CONCLUSION

In this work, we proposed a way to find the best and relevant answers to asked questions from previously asked similar questions where two different algorithms Knn and naïve bayes are compared. We got better results for Knn compared to naïve bayes with results as well as in terms of time required. In this system first similar questions are retrieved for a given question and pool of answers are collected which are given to rank model where it ranks answers based on features which will provide user most relevant answers in less time, also user can rate those answers for further retrieval. In future noisy data can be handled related to multi-topical questions for that work can be done related to topic modeling for general community question-answering to improve answer ranking. Also Query-focused summarization can be implemented to provide summery to new question from historic archives where user will get summery of multiple answers so that one complete and most relevant answer can be formed.

REFERENCES

[1] Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang, "Data-driven Answer Selection in Community QA Systems", *IEEE transactions on knowledge and data engineering*, June 2016

[2] Sahu, T. P., Nagwani, N. K., & Verma, S. "Selecting Best Answer: An Empirical Analysis on Community Question Answering Sites", *IEEE Access, 4, 4797–4808. doi:10.1109/access.*2016.

[3] Calefato, F., Lanubile, F., & Novielli, N." Moving to stack overflow: Best-answer prediction in legacy developer forums". *In Proceedings of the 10th ACM/IEEE international symposium on empirical software engineering and measurement. Article 13(pp. 1–10). ACM,*2016

[4] Y. Tian, P. S. Kochhar, E.-P. Lim, F. Zhu, and D. Lo, ''Predicting best answerers for new questions: An approach leveraging topic modeling and collaborative voting,*'' in Proc. Workshops Int. Conf. Social Informat., 2013, pp. 55–68.*

[5] Z. Ji and B. Wang, "Learning to rank for question routing in community question answering," *in Proceedings of CIKM'13.ACM, 2013, pp.* 2363–2368.

[6] F. Hieber and S. Riezler, "Improved answer ranking in social question-answering portals*," in Proceedings of SMUC'11. ACM, 2011, pp. 19–26*

[7] Guoxin Liu, Haiying Shen," iASK: A Distributed Q&A System Incorporating Social Community and Global Collective Intelligence", *IEEE transaction 2017*

[8] Dalia Elalfy, Walaa Gad, Rasha Ismail, "Predicting Best Answer in Community Questions based on Content and Sentiment Analysis", *IC1CIS'15*

[9] Oleksandr Kolomiyets, Marie-Francine Moens,"A Survey on Question Answering Technology from an Information Retrieval Perspective*", publication in Information Sciences, August 2011*

[10] W. Wei, Z. Ming, L. Nie, G. Li, J. Li, F. Zhu, T. Shang, and C. Luo, "Exploring heterogeneous features for query-focused summarization of categorized community answers," *Inf. Sci., vol. 330, pp. 403–423,* 2016.

[11] Zongcheng Ji, and Bin Wang,"Learning to Rank for Question Routing in Community Question Answering", *ACM* 2013

[12] Show-Jane Yen, Yu-Chieh Wu, Jie-Chi Yang, Yue-Shi Lee,"A support vector machine-based context-ranking model for question answering", *Information Sciences,* October 2012