Health Adviser: Social Question and Answer System using Datamining

Satish Kumar.T, Vishwakiran, Preran Devaiah

Abstract*: Question and answer systems play a crucial role in providing answers to users that seek it, primarily in the case of non-factual questions. A Social Q&A system aims to integrate properties of social networks like similar interests of the users and mutual trust between within a set of friends which makes the reliability of certain answers more prudent. In simple terms, it aims to find a expert that can provide the right answers, ideally within the user's social circle. Our project endeavors to serve pregnant mothers in answering their private or non-private questions regarding their pregnancies. Our system provides a platform for the to-be mothers to ask their questions to experts who can suggest suitable advises for their benefit. The user also has the added advantage of accessing information from experts who belong to their social circle thereby reducing the anonymity of the expert. We have compared the accuracy of few machine learning algorithms that includes Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor kNN), Logistic Regression (LR), ZeroR (the baseline classifier). Our model has an accuracy of over 75%, while demonstrating robustness across learning algorithms.

Index Terms*: Decision tree, Data Mining, Health Adviser, *Social Q&A system.

I. INTRODUCTION

The web as we know of it today consists of multiple sources that can provide answers to a user's queries. The most standout example of this, are popular search engines like Google and Bing. Google uses techniques like keyword matching and tokenization of the query to access the right answers. Bing software was recently licensed by Facebook to deliver to its customers an integrated platform of search engine that involves the social networks of the user. Needless to say, this was a huge move towards the future. However, there lie some shortcomings in this approach. Search engines perform well when question posed to it are factual questions. Suppose a question like "Who is best batsman currently in the world" the system fails to deliver accurate results. This is where social Q&A systems come to the fore. These systems work best when non-factual data is needed. It is also an added advantage that factual questions will anyway be answered as factual as possible as its accuracy is public. Q&A software is often provided to business and technical industries, so its users can be asked questions as well as provides or receive expert answers to them. This kind of software is useful in industries like the above mentioned as newcomers can pose non-formal questions to publically which can be answered by any expert within its system.

II. LITERATURE SURVEY

A thorough understanding of few research papers provided with the ideas to implement this project. Although some applications prefer searching as opposed to asking, generally provides a more comprehensive answer to queries [1].
The scripts give an understanding that teens are more likely to ask factual questions and that too in higher numbers compared to adults [2]. Some resources conducted tests on Question Retrieval processes and denoted ways of integrating users' availability and expertise to acquire the answer within a frame of time [5]. It is noted that some techniques used grouping users into sets of questions answered or posted by them. Discussions that revolved around factual questions often were observed to have a thread of answers. When researching non-factual questions, it is noted that the threads of communication were longer as it most probably lead to further questioning from the user and subsequent answers from other field experts. On an average it could be said that factual questions often yield a smaller thread of communication while non-factual questions have a longer thread due to subsequent information seeking. Few papers on this topic also helped enormously in understanding relationships between user groups which help in categorizing users into pools within which users are most likely to question about a specific domain and receive expert opinion from within the same group [6]. Since this project focused on the medical application of the system, some books gave an research insight that suggested the efficient use of social question and answer systems in the field of medicine [7,8]. PeopleRank was studied, which is a forwarding algorithm exploiting social properties to decrease the number of message retransmissions in Mobile Opportunistic Networks. PeopleRank is a social distributed algorithm which measures opportunistically the significance of a node in a social graph depending on the social interaction between nodes and their interaction frequency.

Revised Manuscript Received on 30 July 2019.

***** Correspondence Author
Satish Kumar .T*, Computer Science and Engineering, BMSIT&M, Bangalore, India. satish.savvy@gmail.com.
Vishwakiran, Computer Science and Engineering, BMSIT&M, Bangalore, India. vishwakiran@bmsit.in.
Preran Devaiah, Research Scholar, Computer Science and Engineering, BMSIT&M, Bangalore, India. prerandevaiah@gmail.com.

It succeeds in establishing an end-to-end delay and a success rate given by flooding, while reducing the percentage of retransmission by 50% [9].

III. SYSTEM ARCHITECTURE

A Social Q&A system aims to integrate properties of social networks like similar interests of the users and mutual trust between within a set of friends which makes the reliability of certain answers more prudent. In simple terms, it aims to find a expert that can provide the right answers, ideally within the user's social circle. Fig 1 displays the project architecture.

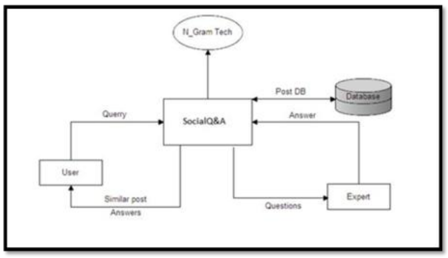

Fig 1: Architecture of The System

The research work aims to provide a system that abides with security and efficiency enhancements by protecting our end user's privacy, identifies, and finds the answers automatically for questions that have been posed earlier. Our results have shown us that Social Q&A systems can be used to ingeniously increase answers quality and fetch time of the answers. In the proposed system, the user interacts with the system through queries that are subject oriented. The system reverts back with previous posts and answers that are similar. When it encounters a query that has not been posed till date, it consults an expert and stores the answer that is provided by the expert in the database for future encounters from the user.

IV. SEQUENCE DIAGRAM

Fig 2 captures the sequence diagram of the project. The user in need of the system proposes the system with the question which is then displayed on the interface. The interface then uses the software to generate a token for the current question asked by the user. Post the token generation operation, an event is triggered to fetch the question from the database. The database fetches the question as per the query posed to it by user.

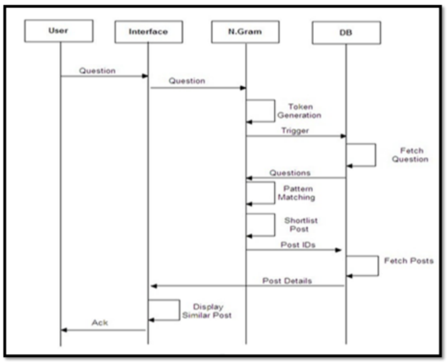

Fig 2: Sequence Diagram Of The Sessions

The fetched question then undergoes the pattern matching task to look for the common keywords to match the question. The next stage involves short listing a few posts that might have the relevant answer to the query posed. If there are posts of similar type, the post id is matches within the database of our system and similar posts are retrieved hoping that these will answer the users query. Once we have a confirmed id we then fetch the post from the database and then send it to the interface via the software which is then provided to the end user.

V. IMPLEMENTATION OF THE PROJECT

The framework that is developed is a question and answer system and takes into consideration the social relation of the user and the person answering their question. This project on is specifically build considering pregnancy as the domain. We hope to provide a suitable system for a pregnant mother to pose her questions and receive answers from trusted sources. We have implemented this project using Java as the language of code. We considered the usability of the language and its universal popularity for this consideration, which would make future enhancements easy configure. Initially, we operate by classifying users into groups of similar questions posed/ answers needed. This gives us an understanding of what the user wants and most importantly helps us retrieve any past data that might be similar to the ones that are needed right now. We use the principle of Bloom filters to assess if the user is within a given set of grouping. Bloom filter is a probabilistic data structure that is used to test if a member is present in the set or not. The price paid for this efficiency is that it tells us that the element is not in the set or may be in the set. To ensure that there is privacy offered to the user, we also consider Onion Routing which is a technique for anonymous communication over a network. Like in an onion, the message is encapsulated in multiple layers of encryption. In the interface, we have three modules of functionality- User, Admin and Expert. The user poses the question that he/she seeks the answer to. The admin can control which user to allow or which expert to be added to the system. If the user posts a question that is not posed before, the system contacts the expert to get an opinion. This opinion is then saved for future references. The answers can be posted on the home page, to which replies will be generated publically, or the questions can be asked in private, which will yield private answers to the users. Researchers have applied several statistical classification techniques for text categorization. In this research we use five different classifiers including, SVM, DT, LR, ZeroR and kNN, which have been shown to be elective in previous text classification work. The data space is partitioned using linear or non-linear boundaries between different classes by SVM. By partitioning the data space SVMs achieves high performance in text categorization, this is because they accept high dimensional feature spaces and sparse feature vectors[11]. No the other hand the KNN by considering its neighbors based on a similarity measure (e.g., distance functions) classifies new text by a majority vote of its neighbors.

The KNN algorithm is simple and fast, also it is sensitive to the local structure of the data. The next algorithm for classification problems is called the Zero Rule algorithm, also called ZeroR. For a regression predictive modeling problem where a numeric value is predicted, the ZeroR algorithm predicts the mean of the training dataset. The other famously used machine learning algorithm is the Logistic regression. It many ways linear regression and logistic regression are similar. Linear regression is usually used to predict/forecast values but logistic regression is used for classifying tasks. By using DT classifier, on the other hand represents leaves as class labels and branches as conjunctions of features that lead to those class labels. It is observed that DT trees are consistently slow and sometimes suffer from over-fitting. Decision Tree: Decision tree is one of the most useful and technically strong tools for classification and prediction. A Decision tree is a flowchart like structure, where each internal node represents a decision criteria/test, each branch emanating from it shows the test result and leaf nodes are class labels. Decision trees are based on the concept of recursive partitioning till the class labels consist of identical entries.

In our implementation, we use decision trees to classify the user's query. Each query is classified as one of three categories – Database, Social Network or Searching. The classifier works on principles of tokenization of the query, searching the social network for experts and pattern matching. If the classifier denotes the query as Database, the token of the queries used as the primary key to obtain previous instances of answers to the same question within the database. On the other hand, if will be classified as Social, the classifier has noted that there is no instance of previous answers but the friend circle might contain a expert who can provide the answer. Here, the social circle of the user is given preference. Any experts within the social system have the opportunity to provide answers to the posed query. If the social circle does not consist of experts, the classifier denotes it as Searching. Here, the query is posted publically to all experts to obtain suitable answers.

Psuedocode for decision tree implementation:

Input: A set of users and experts, by the admin, after authorization

```
1. While the user is valid,
    a. For each query,
    Apply the classifier,
    The classified queries will be of three types : Database, Social or Searching
2. If query=database,
    a. Obtain the token or primary key
    b. Match the token with existing entries
    c. Retrieve the answers
3. If query=social,
    a. Post the question to experts in social circle
    b. Retrieve answers
    c. Store answers in database with primary key as token of query
4. If query=searching,
    a. Post the question to experts publically
    b. Retrieve answers
    c. Store answers in database with primary key as token of query
```

The graph gives the accuracy values of all the algorithms used in this research, it is clearly evident that the DT algorithm proves to be better than any algorithms considered in this experiment.

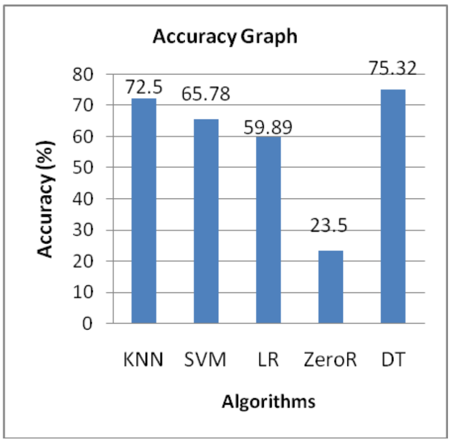

Fig 3: Bar graph representing the Accuracy of various Algorithms.

VI. CONCLUSION AND FUTURE ENHANCEMENTS

Q&A system are used in various industries for applications such as research activities, information gather and retrieval and open forum talks or discussions. The aim of this system that we have developed is to integrate ways of improving the access time of answers within the system and the quality of answers that a user receives. In this regard, we have developed a Social Q&A system for application in the field of pregnancy related queries.

It takes advantage of multiple social network properties, to forward a question to potential answer providing experts, ensuring that a posed question receives a high-quality answer within a period of time that we feel is appropriate. It also provides an added advantage of experts receiving questions on their homepage which is easier than the earlier practice of having to search a search engine to find questions that he/she can find answerable. We have tried to gain inspiration and tried to implement principles of Bloom filters and Onion Routing. Bloom Filter is a probabilistic method which allots a probability to the chances of a user belonging to a particular subset. Onion Routing is a routing protocol where, the message that passes through each layer of communication gets added with a layer of encryption. This yield a model where the message is layer-on-layer stacked with encryptions which we have tried to implement in our private query postings.

Q&A activities from our system have shown that our application provides answers with minimal latency and also provides answers of high quality. Compared to algorithms like SVM, kNN, LR, and ZeroR, DT gives a accuracy of more then 75% to the answers that are predicted.In the future, we hope to include techniques of topic modeling and embedding of words to ensure the efficiency of the system can be increased.

The dynamism of Social Q&A system also can be used to integrate it with machine learning techniques to ensure the answers can be learned by the system and innately researched and displayed to the user, even n cases of prior history of the question being non-existent. We will conduct tests on a large user base in the real-world experiment.

REFERENCES

1. M. R. Morris, J. Teevan, and K. Panovich. A Comparison of Information Seeking Using Search Engines and Social Networks. In In Proc. of ICWSM, 2010.
2. M. R. Morris, J. Teevan, and K. Panovich. What do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Q&A Behavior. In Proc. of CHI, 2010.
3. Gyongyi, G. Koutrika, J. Pedersen, and H. Garcia-Molina. Questioning Yahoo! Answers. In Proc. of QAWeb, 2008.
4. Yahoo!Answers Team. Yahoo! Answers BLOG. http://yahooanswers.tumblr.com/, [Accessed on 10/20/2014].
5. B. Li and I. King. Routing Questions to Appropriate Answerers in Community Question Answering Services. In Proc. of CIKM, 2010.
6. L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In Proc. of WWW, 2008.
7. G. Drosatos, P. Efraimidis, A. Arampatzis, G. Stamatelatos, and I. Athanasiadis. Pythia: A privacy-enhanced personalized contextual suggestion system for tourism. In COMPSAC, 2015.
8. S. Li, Q. Jin, X. Jiang, and J. Park. Frontier and Future Development of Information Technology in Medicine and Education: ITME 2013. Springer Science & Business Media, 2013.
9. A. Mtibaa, M. May, C. Diot, and M. Ammar. Peoplerank: Social Opportunistic Forwarding. In Proc. of Infocom, 2010.
10. E. Pennisi. How Did Cooperative Behavior Evolve? Science, 2005.
11. Thorsten Joachims, “Text categorization with support vector machines: Learning with many relevant features," in Proceedings of ECML-98, 10th European Conference on Machine Learning. 1998, pp. 137-142, Springer Verlag, Heidelberg, DE.

AUTHORS PROFILE

Dr. Satish Kumar.T holds B.E from Bangalore University, M.Tech and Ph.D in Computer Science & Engineering from ANNA University.
He is currently working with Department of CSE, BMSIT. Prior to this he was associated with RNS Institute of Technology, Bangalore. He has 19 years of professional experience, which spans from Industry, academics, research and consultancy. He has published around 20 papers in reputed International Journals / Conferences. His research primarily focuses on Code Optimization on High performance Computing (HPC) systems using heuristics methods. Specifically, using Multi-core clusters with high degree of computations. Over the years Code optimization, while highly relevant to HPC is slowly picking up grounds in mobile computing and IOT. His current research extends to development of Compiler Optimization Algorithms, design of RTOS and Embedded system design, and its synchronization. He is a member of Indian Society for Technical Education and IEEE. He was also the BOE member of VTU, Belagavi and journal reviewer for several journals.

Dr. Vishwa Kiran is an Assistant Professor at BMS Institute of Technology and Management. He has 17 years of experience in embedded software development and training. He is a consultant for Pushkala Technologies Pvt Ltd and Aprameyah Technologies Pvt Ltd for development and corporate training activities. He has handled a corporate training for Texas Instruments, Cisco, Siemens, Western Digital, L&T Infotech, Nokia, Samsung, Sasken, Wipro, Optis Information Services and Infinite Solutions. He was a full-time Ph.D. Scholar between November 2013 to October 2016 at University Visvesvaraya College of Engineering. He has presented his research work in various conferences like INDICON-PUNE(2014), ICIIS - GWALIOR (2014) and TENCON-Macau, China(2015). His research work is published in Springer and Inderscience Journals. He is passionate to work on Linux driver development and building Android for embedded systems. He received a Bachelor of Engineering degree in Electronics and Communication from Bangalore University, Master of Technology in Computer Science and Engineering from Visvesvaraya Technological University and Ph.D. in Computer Science and Engineering from Bangalore University.

Preran Devaiah has a great passion for the Computer Science and Data Analytics and Data Science. He did under graduation in BE Computer Science programme from BMSIT & M college under VTU University.
His achievements include representing state in Basketball in Pre-University. His aim to give back to the society prompted him to become a Polling officer for the State Assembly Elections and Blood Drive donations in college. Academically, he was under the Prime Minister scholarship for Ex-Servicemen’s Children for 2 years of his engineering degree.

.

Retrieval Number: B2802078219/19©BEIESP
DOI: 10.35940/ijrte.B2802.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication