International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 4 | ISSN : 2456-3307

Generating Semantic Frames from Queries Using Natural Language Query Processing In Concertize Search Engine

[1]Pranali Khambalkar, [1]Prachi Sawwalakhe, [1]Nandini Kewale, [1]Rutuja Margadhe, [1]Samiksha Sahare, [1]Anshul Rahule, [2]Prof. Umesh S. Samarth

[1]BE Scholars, Department Of Computer Science & Engineering, J.D. College of Engineering and Management, Nagpur, Maharashtra, India

[2]Assistant Professor, Department Of Computer Science & Engineering, J.D. College of Engineering and Management, Nagpur, Maharashtra, India

ABSTRACT

Question and Answering (QA) frameworks are alluded as virtual associates and are imagined to be the cutting edge call focus. However the precision of such QA frameworks isn't as attractive and necessities critical upgrade. Understanding the goal of the inquiry is a critical supporter of an effective framework which has not been frequently broke down. To answer English dialect questions utilizing PC is an intriguing and testing issue. For the most part such issues are dealt with under two classifications: open Domain issues and close space issues. This paper exhibits a framework that endeavours to explain both close and open area issues. Answers to inquiries from close area can't be looked utilizing a web index. Henceforth answers must be put away in a database by an area master. At that point, the test is to comprehend the English dialect question with the goal that the arrangement could be coordinated to the separate answer in the database. We utilize a format coordinating method to play out this coordinating. The framework is produced with the end goal that the inquiries can be asked utilizing short messages from a cell phone and accordingly the framework is intended to comprehend SMS dialect notwithstanding English.

Keywords— Answering System, Template Matching, Natural Language Processing, QA System; NLP Algorithms; Semantic Similarity; User Intent

I. INTRODUCTION

Evolution of web from a read only to read write mode has made way for a huge load of information in the form of knowledge bases. Wikipedia, Freebase, YAGO, Microsoft Satori and Google Knowledge Graph are some of the well-known knowledge bases [1]. Information present in them could be used to build specific decision making /advisory systems. QA systems, which are a part of advisory systems are viewed as futuristic replacement of call centers and are called as virtual assistants. QA systems generally are classified based on the type of queries asked by the user and by the way system retrieves information while responding to the queries. While the former could be again classified as supervised (frequently asked questions (FAQ)) and unsupervised (generic questions) the latter could be classified based on logical reasoning, semantic understanding or plain key word matching.

Most of the systems reported in literature are FAQ/key word matching type while semantic /logical reasoning systems have been rare. Common objective of all QA systems has been to find a relevant response to a precise natural language query. In spite of the recent advances, accuracy and performance are the two cardinal areas in search query processing where there is still a huge scope for enhancement. Accuracy issues could be attributed to the fact that the typical queries are generally insufficient and don't completely describe the user's need. Hence to classify the huge content into predefined categories presents a huge challenge.

Literature reports use of machine learning algorithms to train a classifier and predict the category of an input query. However accuracy of such systems could be enhanced only when both discriminative features as well as sufficient sample size co exists, which is a rarity in a real world scenario. It must be noted that an ideal system should be context aware and be able to respond to the queries with high accuracy. Hence understanding the intent of the user is important for providing relevant responses to the user queries. Another significant factor that has to be taken care of is the ever growing size of the content. Optimal method of indexing the content and scaling the solution is also as important as the response of these systems. However with recent advances in cloud and distributed computing the scalability part could be solved.

QA systems have evolved from a very generic solution provider to be more specific to a particular domain. Healthcare and retail are the domains that have started to deploy these systems [16-18]. Primary objective of the current study is to develop a context aware QA system using an improved approach that would be able to provide relevant responses using algorithms in supervised and unsupervised model followed by a novel scoring mechanism.

II. Literature Review

Link analysis in social media. In particular, link-based ranking algorithms that were triumphant in estimating the quality of web pages have been applied in this context. Link based methods have been shown to be triumphant for several tasks in social media [2].Two of the most prominent link based ranking algorithms are Page Rank[4] and HITS [3] Consider a graph G = (V;E) with vertex set V with respect to the users of a question/answer system and contains a directed edge e = (u; v) є E from a user u є V to a user v є V if user u has answered to at least one question of user v.

Expertise Rank [8] respect to Page Rank over the transposed graph G′= (V;E′), that is, a score is cultivated from the person receiving the answer to the person giving the answer. The recursion implies that if person u was able to give an answer to person v, and person v was able to provide an answer to person w, then u should receive some extra points given that he/she was able to provide an answer to a person with a certain degree of expertise.

Propagating reputation. Guha et al. [5] study the problem of propagating trust and distrust among Epinions users, who may assign positive (trust) and negative (distrust) ratings to each other. The authors study ways of integrating trust and distrust and observe that, while appraising trust as a transitive property makes sense, distrust cannot be considered transitive.

Ziegler and Lausen [7] has also studied the models for propagation of trust. They present an anatomy of trust metrics and discuss ways of incorporating information about distrust into the rating scores.

Question/answering portals and forums. The particular framework of question/answering communities we focus on in this paper has been the object of some study in recent years.

According to Su et al. [6], the quality of answers in question/answering portals is good on average, but the quality of specific answers varies considerably. In particular, in a study of the answers to a set of questions in Yahoo! Answers, the authors found that the fragment of correct answers to specific questions asked by the authors of the study, varied from 17% to 45%. The fragment of questions in their sample with atleast one good answer was much higher, varying from 65% to 90%, meaning that a method for finding high-quality answers can have a remarkable effect in the user's satisfaction with the system.

Expert Finding. Zhang et al. [8] examine data from an online forum, seeking to identify users with high expertise. They research the user answers graph in which there is association between users u and v if u answers a question by v, pertaining both Expertise Rank and HITS to identify users with high expertise. Their results gives high interrelationship between link-based metrics and the answer quality. The authors also develop synthetic models that record some of the characteristics of the interactions among users in their dataset.

The HITS algorithm is run on the user-answer graph. Jurczyk and Agichtein [9] presents an application of the HITS algorithm [3] to a question/answering portal. The results finds that HITS is a promising approach, as the obtained authority score is better interrelated with the number of votes that the items receive, than simply counting the number of answers the answerer has given in the past. Dom et al. [10] studied the interpretation of several link-based algorithms to rank people by expertise on a network of e-mail exchanges.

Text analysis for content quality. Most work on evaluating the quality of text has been in the field of Automated Essay Grading (AES), where writings of students are sorted by machines on several aspects, including compositionality, style, precision, and soundness. AES systems are typically shown to correlate very well with human judgments. Although simplistic and disputable, these methods are widely-used and provide a rough estimation of the difficulty of text.

Implicit feedback for ranking. Implicit feedback from millions of web users has been shown to be very important source of result quality and ranking information. Authors had incorporate the results on click interpretation on web search results from these studies, as a source of quality information in social media. In particular, clicks on results and methods for interpreting the clicks have been studied in references [11].

III. Implementation

In the related study, we ran over numerous courses by which the substance quality can be assessed in the inquiry noting entrance based on inputs enrolled by various clients. Yet, the issue associated with such a plan is, to the point that the clients input is compulsory in finding the evaluations to the solution so discovering importance among the appropriate responses Here is an endeavour to acquire the most applicable answers among all the given answers without's client criticism enlistment.

For this reason, as we are managing the group driven inquiry noting entryway where for various inquiry and answers can be given so we are utilizing the idea of computing the scores for every one of the appropriate responses given for an inquiry. In the wake of having scores for every one of the appropriate responses and scores can be created by applying some component to it will be clarified letter. Presently it is anything but difficult to gauge the astounding substance among every one of the appropriate responses by taking the high scorer reply on the best. There are a few parameters on which the assignment will achieve.

QA framework that has been produced is an intranet arrangement. The client can ask questions either by composing utilizing the pursuit enclose the UI or through a voice input. Google API was utilized to change over the voice contribution to content and play out the vital tasks on the inquiry. The framework had two modes in particular managed and unsupervised, both of which have been clarified. Knowledge base was built by slithering FAQ open sites of insurance agencies and put away in various level records. All the conceivable inquiries were marked with help of a specialist. Reactions to the client questions depended on watchword coordinating.

The general engineering of the framework can be subdivided into three primary modules: (1) Pre-handling, (2) Answer Discovery, and (3) answering. Figure 1 demonstrates the framework design of the inquiry and noting framework.

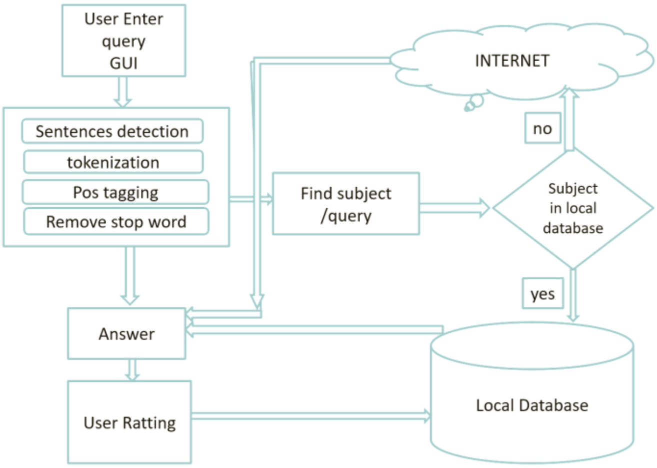

Figure 1 System Architecture

A. Pre-Processing Module

Pre-Processing module for the most part comprises of two activities:

- Recognizing the sentence and identify the catchphrases.
- Expelling stop words and stemwords.

The framework is relied upon to process writings with both English and SMS dialects it is important to supplant the SMS shortenings with the comparing English words previously handling client addresses further. This is finished by alluding to pre-put away habitually utilized SMS truncations. Stop words and Stem words are the words that add no impact to the significance of a sentence, are evacuated. Stemming is a pre-preparing venture in Text Mining applications and in addition an exceptionally basic necessity of Natural Language handling capacities. Truth be told it is essential in the majority of the Information Retrieval frameworks. The principle reason for stemming is to diminish distinctive linguistic structures/word types of a word like its thing, descriptor, verb, intensifier and so forth to its root shape. We can state that the objective of stemming is to decrease inflectional structures and at times derivationally related types of a word to a typical base shape. Evacuating stop words and [5] [6] stem words is done to build the viability of the framework by sparing time and plate space. We are utilizing Porter stemming calculation for stemming reason. Pre-Processing is done to have enhanced coordinating format of client made inquiry.

B. Answer Discovering Module

- Question-Template coordinating module

The pre-handled content is coordinated against every last pre put away layout until the point that it finds the best coordinated format with the got content. So as to do this, layouts are made by a particular linguistic structure. Facilitate in this module, words that are considered to have equivalent words are alluded in an equivalent word document. This equivalent word record can be adjusted by the significant area and are refreshed from a standard database. It is important that the formats here are for questions and not for answers. The primary focus of this framework is to distinguish the nearest format that matches the inquiry we have gotten from the client.

- Web-Data Extraction Module

The framework is produced in such way that if client made inquiry is out of close area at that point internet searcher will scan for suitable answer and afterward framework will restore the precise response to the end client. Numerous destinations now bolster APIs that empower PC projects to gather data. A few Web-Scrapping arrangements are accessible. For web information extraction we are utilizing JSON which make an interpretation of HTML into other organization and makes it less difficult to remove the coveted substance. JSON is an all-inclusive, dialect autonomous configuration for information. It depends on object literal documentation of JavaScript.

C. Answering Module

Since every single layout speaking to an inquiry are re-established in a database with its answer, exactly when the best coordinated format for the inquiry is found, [2]the comparing answer will be come back to the end client.

As said before, the client questions are addressed utilizing format coordinating. In this area, we talk about the layouts utilized and their language structure. [1]Our strategy depends on physically indicating formats for each Frequently Asked Question. Those are put away in a database combined with the appropriate responses. The formats are coordinated against the inquiries requested that by clients locate the best coordinated layout. The achievement of the inquiry noting in this manner depends a great deal on the nature of these layouts.

[1]The linguistic structure of the formats is characterized with the goal that a solitary layout could coordinate a wide range of variations of a similar inquiry. An inquiry may be asked in various courses because of at least one of the accompanying reasons: distinctive tenses; solitary/plural structures; use of equivalent words; the request of utilizing words; and the utilization of discretionary words. Utilizing the above sentence structure discretionary complex formats can be built. Additionally

expressions can be settled inside each other, and equivalent word rundown could likewise contain phrases that have an indistinguishable significance from a solitary word.

Following are distinguished as the upsides of utilizing a format coordinating methodology: (1) Precision of the recovery is high on the grounds that the catchphrases are chosen utilizing human insight; (2) It is an advancing framework, since its inquiry noting capacity enhances as more inquiries are asked, and new FAQ sections are added to the database; and (3) A comprehension of the issue area isn't required for creating. The fundamental inconvenience of the framework is that the layouts should be composed physically for all questions.[1]The format coordinating method is upgraded utilizing two extra procedures and they are: (1) applying disembowelling and (2) utilizing an equivalent word list. It is trusted that the greater part of the spelling botches happen as a result of oversight, expansion or out of request vowels. In this manner, evacuating vowels in a sentence will decrease the measure of spelling botches experienced in a sentence. Along these lines vowels are expelled from client inquiries in our framework. The way toward evacuating vowels in content is known as disemvoweling. [3]Disembowelling is additionally done in our layouts as a methods for representing spelling botches in client questions and for simple coordinating of the formats. We trust this is crucial expansion to the framework as the framework is relied upon to be utilized by non-local English speakers who are inclined to commit plentiful spelling errors in their inquiries. Rather than straight pursuit, to lessen the coordinating tally between the clients asked inquiry and database put away format.

IV. Conclusions

We presented an friendly automatic answering system with the capacity of identifying and noting questions asked in English or. Exactness is a noteworthy constraint in the greater part of the QA frameworks. Understanding the goal of client could decide the precision of the QA framework reaction. The investigation rovides a novel method for understanding the goal of the question and gives a scoring instrument to distinguish related substance and concentrate significant data from that point for a given query.This application can be utilized by client mind operators to deal with telephonic calls though they don't need to check database physically for answers and are taken care of by this application with fast reaction. In this situation, the importance criticism from the specialist can go about as a contribution to enhance the exactness of the framework. In case of another altered answer, that comparing question - answer match can be nourished to the framework for preparing the FAQ demonstrate.

V. REFERENCES

[1]. Dong, X. L., Murphy, K., Gabrilovich, E., Heitz, G., Horn, W., Lao, N. & Zhang, W. (2014). Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion.

[2]. Carlos Castillo, Debora Donato et al., “FINDING HIGH QUALITY CONTENTS IN SOCIAL MEDIA”Yahoo! Research Barcelona, Spain, WSDM'08, February 11.12, 2008.

[3]. J. P. Scott. Social Network Analysis: A Handbook.SAGE Publications, January 2000.

[4]. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604{632, 1999.

[5]. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[6]. R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In WWW '04: Proceedings of the 13th international conference on World Wide Web, pages 403{412, New York, NY, USA, 2004. ACM Press

[7]. Q. Su, D. Pavlov, J.-H. Chow, and W. C. Baker. Internet-scale collection of human-reviewed data. In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 231{240, New York, NY, USA, 2007. ACM Press.

[8]. C.-N. Ziegler and G. Lausen. Propagation models for trust and distrust in social networks. Information Systems Frontiers, 7(4-5):337{358, December 2005.

[9]. J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In WWW '07: Proceedings of the 16th international conference

on World Wide Web, pages 221{230, New York, NY, USA, 2007. ACM Press.

[10]. P. Jurczyk and E. Agichtein. HITS on question answer portals: an exploration of link analysis for author ranking. In SIGIR (posters). ACM, 2007.

[11]. B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In Proceedings of Workshop on Data Mining and Knowledge Discovery, pages 42{48, San Diego, CA, USA, 2003. ACM Press.

[12]. T. Joachims, L. A. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting click through data as implicit feedback. In SIGIR, pages 154{161, 2005.

[13]. K. Ali and M. Scarr. Robust methodologies for modeling web click distributions. In WWW, pages 511{520, 2007.

[14]. C. Anderson. The Long Tail: Why the Future of Business Is Selling Less of More. Hyperion, July 2006.

[15]. Y. Attali and J. Burstein. Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment, 4(3), February 2006.

[16]. A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39{71, 1996.