International Journal of Advanced Technology in Engineering and Science
Vol. No.09, Issue No. 01, January 2021
www.ijates.com
ijates
ISSN 2348 - 7550

Content Driven Answer Selection: A Survey

Ms. Manisha V Khadse

Assistant.Professor -Department of B.Voc. IT & ITeS

Skill Development Center Savitribai Phule Pune University, Pune,India

manishakhadse151@gmail.com

ABSTRACT

Now a days term-based and pattern-based algorithms are categorized in the discipline of information filtering to define user's information demands from a pool of documents. A basic demand is that the documents in the pool are all of one topic. But the user's demand can be deferent and the documents in the pool may contain different topics. Despite this has been subsequently utilized in the discipline of information filtering and machine learning, its effectiveness in the area of information filtering has not been well discovered. Community-based question and answering techniques have evolved as a constructive way for knowledge dissemination and to obtain correct information. Several CQA (Community Question Answering) services are present at the moment which includes Yahoo! Answers, Quora and Stack Overflow which allow users to ask complicated questions and to answer other user's questions. These services provide user's explicit and self-contained answers rather than lists of web pages or documents. Thus, CQA services have provided a sensible substitute to general purpose web search. However, there is a significant gap between posted questions and budding answers due to the fast increasing of posted questions and the deficiency of an effective way to find potential answers. This study aims to build a ranking based framework that identifies parameters for the best quality answer in a QAsystem.

Index Terms—Information filtering, Relevance ranking, CQA, Yahoo, Google, QA system

I. INTRODUCTION

Data mining is a best way to drop tautological or unnecessary data from a huge pile of document which represent user's demands .Existing information filtering approaches are developed using a term-based calculation. The main advantage of the term-based approach is its efficient computational performance. It becomes more difficult for users to obtain material of interest, to search for specific structured content or to gain an overview of important and relevant material due to increasing amounts of data. In today's internet era number of peoples is searching for informative data on web, but it is not possible that they could get all relevant data in single web page or document. Search engine provides number of web pages as a search result. But because of advancement in data mining and machine learning approaches this problem has given the new solution. These algorithms return query specific information from large dataset of offline documents and provide user with relevant answer [7][9].

Although existing search engines are capable of providing data needs of internet users, but they are not

capable of executing or analyses termed queries which are submitted in the form of NLP questions. So to tackle this problem, question answering techniques has been built which reply to the natural language queries by providing concise answers.

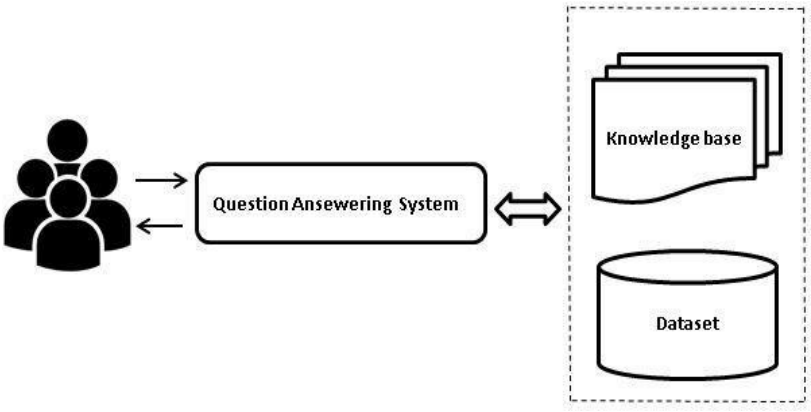

Fig 1.1 QA system

QA system is an important part of information retrieval (IR) domain which directly provide a precise data that satisfies user's information needs which eliminate tedious task of crawling through a list of web pages and documents. As shown in Fig 1.1 QA systems are built on dataset of domain specific knowledge which is a combination of different NLP tools to find best answer.

Main issue in question answering system is identifying parameters for enhancing the standard of the retrieved answer is difficult because deferent user interact with the system differently and so there are variety of correct answers to complicated questions so it is important to analyses the answer extraction and formulation techniques which processes the relevant sentences and extract the foremost answer using more linguistic oriented analysis.

II.REVIEW OFLITERATURE

Author believes that the Community Question Answering CQA system requires semantic gap between questions answer pairs is the main problem in the QA system. It also needs a serious modeling of contextual factor. An attentive deep neural network architecture is proposed in this paper. The architecture is having three layers which are namely; Convolution Neural Network, Long Short Term Memory and Conditional Random Field. The SemEval-2015 CQA dataset is used to develop the experiment.

This paper presented Attentive Neural Network architecture to develop the CQA system. CNN and LSTM components are discussed in this paper. To get better result in CQA system, attentive bi-LSTM is used which is a unit of Recurrent Neural Network[1].

In this paper, author proposed a collaborative learning model for answer selection in question answering, which aims to extract a rich feature presentation for sentence embedding. To overcome the weakness of a single neural network in the distributed sentence representation, author have adopted a parallel structure to combine the sentence embedding's generated by different networks. In addition, the keyword overlap feature learnt by a simple-but-tough-to-beat mechanism, i.e., WR model, is combined with the neural networks. To evaluate the proposed framework, author conducted the experiments on the freely available QA dataset, i.e.,

Insurance QA. The experimental results demonstrate that proposed collaborative learning methods outperform the competitive baselines in terms of well-known evaluation metrics[2].

In paper Data-driven Answer Selection in Community QA Systems authors Zhipeng Gao, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Liqiang Nie presented a framework for answer selection. Here author combined offline and online learning search component. Offline learning component implicitly construct three training sample positive, neutral, and negative by data-driven observations. Then author designed a rank model to assimilate these three types of training samples. Online search component first collect a set of answer candidates by finding its similar questions and then apply the offline learned model to rank the answer candidates by pairwise comparison.

To find the accuracy of proposed system author conducted experiment on one general and one vertical CQA dataset. The advantage of offline learning component system is that it removes the labor intensive and time consuming annotation process [1].

In paper Learning from the Past: Answering New Questions with Past Answers authors Gideon Dror, Yoelle Maarek, Idan Szpektor proposed an automatic question answering system. System uses past answers for reference. Author presented question answering algorithm in two stages first stage single out the determined previous questions which is resemble to a new targeted question. In second stage statistical classifier is applied in order to verify whether the current past answer meets the new question needs. For result calculation author evaluated algorithm on an offline annotated dataset, this shows that it is possible to extract high precision answer. Author analyzed the performance of algorithm by constructing a live system that operates three robots, Jane, Alice, Lilly, who act as real users on Yahoo! Answers and provide the answer, when confident enough, live questions in three active categories. Finally author stated that the analysis showed that these three robots perform better than the average answerer[2].

In paper A Predictive Framework for Retrieving the Best Answer authors, Dion Hoe-Lian Goh, Alton Y.K. Chua presented a hierarchical structure that identifies attributes that are responsible for enhancing the standard of the answer in a question answering system. The system performance is carried out using real data collected from Yahoo! Answers. Author stated that the analysis indicate that the quality of foremost answer is affected by the textual attributes instead of non-textual attributes. Completeness, precision, and reasonableness of the answer signify textual attributes. stature and authenticity of the answerer and asker or the textual attributes like length and language signify non-textual attributes. Author believes that using proposed approach, it will be easy to perceive the significant attributes in expanding the answer extraction systems for the retrieval of better quality answers in a question answering system[3].

In paper JAIST: Combining multiple features for Answer Selection in Community Question Answering authors Quan Hung Tran1 , Vu Duc Tran1 , Tu Thanh , Minh Le Nguyen1, Son Bao proposed an answer ranking based approach to categorize answers in Community QA system. System comprises of three different steps.

Preprocessing: Steps are achieved using The Stanford Core Natural Language Processing Tool.

Building models from data: Here, author described the components and identified attributes that are used, or built for extracting attributes.

Word-matching feature group: This step uses the term based resemblance between the answer and the

question to generate score. Finally author stated that the proposed system generates high results in the task, but there are still some loopholes in the system[4].

III. SYSTEM ARCHITECTURE / SYSTEMOVERVIEW

The below fig 3.1 depicts the flow of the system architecture. Proposed QA system Comprises of three modules. These three main modules are-

1) Query Processing module
2) Document Processing module (information retrieval)
3) Answer Processing module

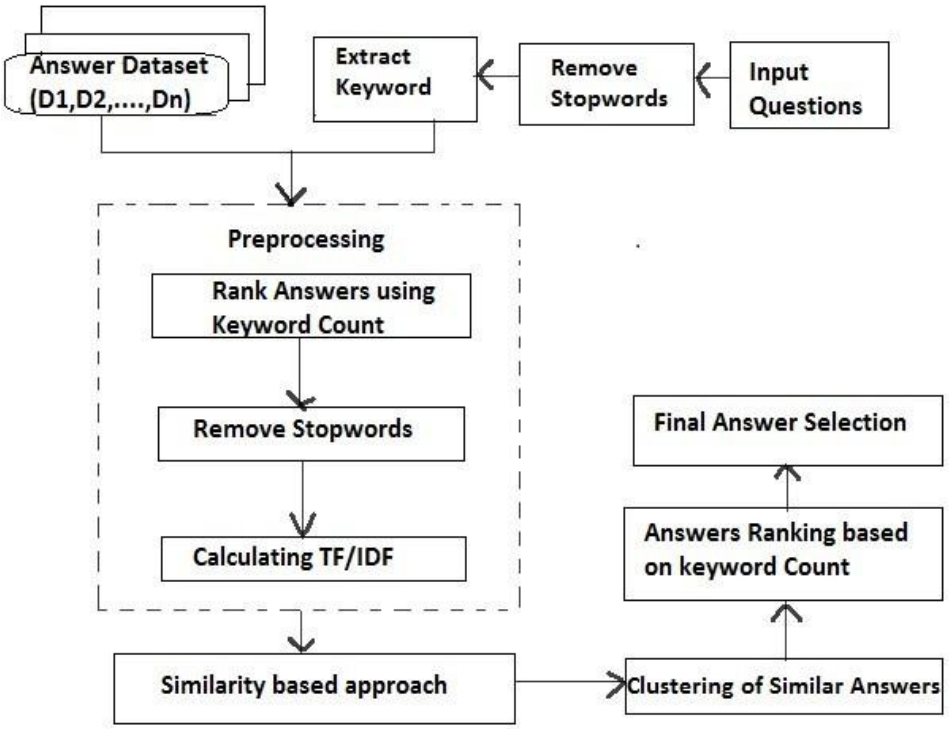

Fig 3.1.Block Diagram of Proposed System

Query processing module: processing of input questions and selected multiple answers. Tokenization: It breaks the sentence into separate lexical words that are separated by white spaces or comma

Stop Word Removal:

Stop words removal filter sentences by removing unnecessary stop words which is helpful in keyword searching.

Stemming Suffixes:

Here suffixes are removed for topic detection Example

Sings and Singing - Sing where “s” and “ing” are suffixes added to topic Sing need to be removed for accuracy purpose.

Document processing module: Here we used similarity based approach which identifies the relation between questions and answers and brings set of candidate paragraphs containing answers

Cosine Similarity Approach:

Cosine Similarity measures the similarity between two sentences in terms of the value within the range of [-1, 1]. Cosine similarity is based on basic fundamentals o term based extraction. Term frequency identifies the documents containing the relevantterms

TF (term, document) = Frequency of term / No of Document

$$tf_i = \frac{n_i}{\sum_k n_k} \quad (1)$$

IDF (inverse document frequency) calculates whether the term is rare or common in all documents. IDF (term, document) is calculated by dividing total number of documents by the number of documents containing provided term and taking log ofthat.

IDF (term, no of answers) = log (Total No of answers / No of answers containing term)

$$idf_i = \log \frac{D}{|\{ d : t \in d \}|} \quad (2)$$

TF-IDF is the multiple of the value of TF and IDF for a provided term. The increasing value of TF-IDF is directly proportional to the number of occurrences o term within an answer and with rarity of the term across the corps

$$TFIDF = TF * IDF \quad (3)$$

Answer processing module: Answer processing module extracts and validates the answers from set of generated sentences which are processed by document processing module

Here we are using NEWSUM Algorithm which finds the exact relation between question and answers by choosing phrase or words according to question and classify them into different clusters and give a highest score sentences.

Clustering:

In this module sentences having similar features are clustered together. A hashtable is generated which contains the sentences with the similarity sore, this hashtable is then sorted in ascending order and mean value is calculated from similarity score. Sentences having value greater than mean value is stored in first cluster and sentences having value lower than mean value is stored in second cluster

IV. CONCLUSION

The work related to the topic under study is based on Pattern based content driven answer selection to make the community questionnaire system better. Basic framework for term selections and sentence filtering studied and compared in depth. The analysis of all major manifestations like terms, phrases and relation between the sentences are used to evaluate the efficiency of the system. The evaluations were performed on the basis of topic modeling methods, pattern mining methods and term-based methods.

The parameters like system accuracy and precision can be used to evaluate the performance of the system. Automatic question answering system uses cosine similarity and ranking algorithm to obtain output in minimal time.

REFERENCES

[1] Mohini Wakchaure, Prakash Kulkarni Department of Computer Science & Engineering Walchand college of engineering, Sangli Sangli, India "A Scheme of Answer Selection In CommunityQuestion Answering Using Machine Learning Techniques" Proceedings of the International Conference on Intelligent Computing and Control Systems IEEE (ICICCS2019)

[2] TAIHUA SHAO , XIAOYAN KUI , PENGFEI ZHANG, AND HONGHUI CHEN 1Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China "Collaborative Learning for Answer Selection in Question Answering" 2169-3536 2018IEEE.

[3] Data-driven Answer Selection in Community QA Systems Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING JUNE2016

[4] Learning from the Past: Answering New Questions with Past Answers Anna Shtok Faculty of Industrial Engineering and Management Technion, Israel Institute of Technology Haifa 32000, Israel annabel@techunix.technion.ac.il Gideon Dror, Yoelle Maarek,Idan Szpektor Yahoo! Research MATAM, Haifa 31905, Israel gideondr, dan@yahoo-inc.com, yoelle@ymail.com WWW 2012 –Session: Leveraging User-Generated Content

[5] A Predictive Framework for Retrieving the Best Answer, Mohan John Blooma, Alton Y.K. Chua, Dion Hoe-Lian Goh Division of Information Studies Wee Kim Wee School of Communication & Information Nanyang TechnologicalUniversity

[6] JAIST: Combining multiple features for Answer Selection in Community Question Answering Quan Hung Tran1 , Vu Duc Tran1 , Tu Thanh Vu2 , Minh Le Nguyen1 , Son Bao Pham2 1 Japan Advanced Institute of Science and Technology 2 University of Engineering and Technology, Vietnam National University, Hanoi Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval2015

[7] A Classification-based Approach to Question Routing in Community Question Answering Tom Chao Zhou1, Michael R. Lyu1, Irwin King1,2 1Department of Computer Science and Engineering 2AT&T Labs Research The Chinese University of Hong Kong 201 Mission Street Shatin, N.T., Hong Kong San Francisco, CA, USA {czhou, lyu, king}@cse.cuhk.edu.hkirwin@research.att.com

[8] Quantitative Evaluation of Passag e Retrieval Algorithms for Question Answering Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton MIT Artificial Intelligence Laboratory Cambridge, Massachusetts, USA SIGIR'03, July 28–August 1, 2003, Toronto, Canada. Copyright 2003 ACM1-58113-646-3/03/0007

[9] Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media Jiang Bian College of Computing Georgia Institute of Technology Atlanta, GA 30332 jbian@cc.gatech.edu Yandong Liu Math & Computer Science Department Emory University Atlanta, GA30332yliu49@emory.edu Eugene Agichtein Math & Computer Science Department Emory University Atlanta, GA 30332 eugene@mathcs.emory.edu Hongyuan Zha College of Computing Georgia Institute of Technology Atlanta, GA 30332 zha@cc.gatech.edu April 21-25, 2008 ·

Beijing,China

[10] Adapting Ranking SVM to Document Retrieval Yunbo CAO1, Jun XU , Tie-Yan LIU, Hang LI, Yalou HUANG, Hsiao-Wuen HON Microsoft Research Asia, No.49 Zhichun Road, Haidian District Beijing, China, 100080 {yunbo.cao, tyliu, hangli, hon}@microsoft.com College of Software, Nankai University, No.94 Weijin Road, Nankai District Tianjin, China, 300071 nkxj@hotmail.com, yellow@nankai.edu.cn Copyright 2006ACM

[11] Rank-GeoFM: A Ranking based Geographical Factorization Method for Point of Interest Recommendation Xutao Li1 Gao Cong 1 Xiao-Li Li 2 Tuan-Anh Nguyen Pham1 Shonali Kr ishnaswamy2 1 School of Computer Engineer ing, Nanyang Technological University, Singapore.{lixutao@, gaocong@, pham0070@e.}ntu.edu.sg 2 Institute for Infocomm Research(I2R), A∗STAR, Singapore.{xlli, spkr ishna}@i2r.a-star.edu.sg

[12] Instance-Based Question Answering:A Data-Driven Approach Lucian Vlad Lita Carnegie Mellon University llita@cs.cmu.edu Jaime Carbonell Carnegie Mellon Universityjgc@cs.cmu.edu

[13] Cimiano, P., Haase, P., Sure, Y., & Volker, J. Question Answering on Top of the BT Digital Library. WWW 2006, Edinburgh, UK(2006).

[14] Cody, C. T., Kwok, O. E., & Daniel, S. W. Scaling Question Answering to the Web. Tenth World Wide Web Conference. Hong Kong, China(2001)

[15] Fushman, D. D. & Lin, J. Answer Extraction, semantic clustering, & extractive summarization for clinical question answering. In the Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney(2006).

[16] Hao, T., Zeng, Q., & Wenyin, L. Semantic Pattern for User-Interactive Question Answering. Second International Conference on Semantics, Knowledge, and Grid (SKG'06).(2006)17