Review Paper on Answers Selection and Recommendation in Community Question Answers System

Ms.Kamble Rajani Ravaso
Department of technology
Shivaji university,Kolhapur
Kolhapur,India
e-mail: veeraval121212@gmail.com

Mr. Chetan J. Awati
Department of technology
Shivaji university,Kolhapur
Kolhapur,Indiav
e-mail:chetan.awati@gmail.com

Ms. Sonam Kharade
DeltaTech India
Kolhapur, India
e-mail: skh9624@gmail.com

Abstract—— Nowadays, question answering system is more convenient for the users, users ask question online and then they will get the answer of that question, but as browsing is primary need for each an individual, the number of users ask question and system will provide answer but the computation time increased as well as waiting time increased and same type of questions are asked by different users, system need to give same answers repeatedly to different users.

To avoid this we propose PLANE technique which may quantitatively rank answer candidates from the relevant question pool. If users ask any question, then system provide answers in ranking form, then system recommend highest rank answer to the user. We proposing expert recommendation system, an expert will provide answer of the question which is asked by the user and we also implement sentence level clustering technique in which a single question have multiple answers, system provide most suitable answer to the question which is asked by the user.

***Keywords**- Community-based question answering, answer selection, observation-guided training set construction*

*****

I. INTRODUCTION

As our project purely based on data mining, The data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning , statistics, and database system. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and management aspects, data pre-processing and interface considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

Community Question Answering (cQA) is gaining popularity online. They are seldom moderated, rather open, and thus they have few restrictions, if any, on who can post and who can answer a question. On the positive side, this means that one can freely ask any question and expect some good, honest answers. On the negative side, it takes effort to go through all possible answers and to make sense of them.

For, example, it is not unusual for a question to have hundreds of answers, which makes it very time consuming to the user to inspect and to winnow. The challenge to propose may help automate the process of finding good answers to new questions in a community created discussion forum(e.g., by retrieving similar questions in the forum and identifying the posts in the answer threads of those questions that answer the question well). The accomplishment of cQA and active user participation, question starvation wide exists in cQA forums that refer to the subsequent types. As the number of data seekers on cQA is increased, the waiting time is extended to get answers of their question, because of waiting time users get disappointed. Expert need to answer of all question even if the question is same, i.e. if a question is of same type but still expert need to answer of all question. Given a matter, rather than naively selecting the most effective answer from the foremost relevant question, during this paper, to gift a completely unique Pairwise Learning to rank model, nicknamed PLANE, which may quantitatively rank answer candidates from the relevant question pool. We gift a completely unique approach to constructing the positive, neutral, and negative coaching samples in terms of preference pairs. This greatly saves the long and labour-intensive labeling method. We tend to propose a pairwise learning to rank model for answer choice in cQA systems. It seamlessly integrates hinge loss, regularization, associate degreed an additive terms at intervals a unified framework is totally different from the standard pairwise learning to rank models, and ours incorporates the neutral coaching samples and learns the discriminative options.

II. LITERATURE SURVEY

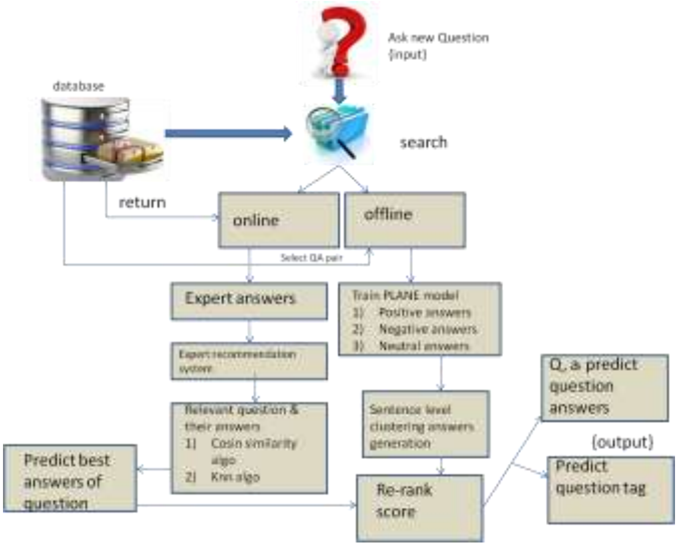
Data-Driven Answer Selection in Community QA Systems”, Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang, To present a novel scheme to rank answer candidates via pairwise comparisons. In particular, it consists of one offline learning component and one online search component. In the offline learning component, firstly automatic establish the positive, negative, and neutral training samples in terms of preference pairs guided by our data-driven observations. We then present a novel model to jointly incorporate these three types of training samples. The closed-form solution of this model is derived. In the online search component, then first collect a pool of answer candidates for the given question via finding its similar questions. "Disease inference from health-related questions via sparse deep learning" L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T. S. Chua, [2] To build a disease inference scheme that is able to automatically infer the possible Diseases of the given questions in community-based health services. First report a user study on the information needs of health seekers in terms of questions and then select those that ask for possible diseases of their manifested symptoms for further analytic. Then next propose a novel deep learning scheme to infer the possible diseases given the questions of health seekers. The proposed scheme comprises of two key components. "MultiVCRank with applications to image retrieval", X. Li, Y. Ye, and M. K. Ng, [22] propose and develop a multi-visual-concept ranking (MultiVCRank) scheme for image retrieval. The key idea is that an image can be represented by several visual concepts, and a hypergraph is built based on visual concepts as hyperedges, where each edge contains images as vertices to share a specific visual concept. "Beyond text QA: Multimedia answer generation by harvesting Web information", L. Nie, M. Wang, Y. Gao, Z. Zha, and T. Chua, [13] To propose a scheme that is able to enrich textual answers in cQA with appropriate media data. Our scheme consists of three components: answer medium selection, query generation for multimedia search, and multimedia data selection and presentation. This approach automatically determines which type of media information should be added for a textual answer. It then automatically collects data from the web to enrich the answer. "A ranking approach on large-scale graph with multidimensional heterogeneous information," IEEE Trans. W. Wei, B. Gao, T. Liu, T. Wang, G. Li, and H. Li, [33] address the large-scale graph-based ranking problem and focus on how to effectively exploit rich heterogeneous information of the graph to improve the ranking performance. Specifically, propose an innovative and effective semi-supervised Page Rank (SSP) approach to parameterize the derived information within a unified semi-supervised learning framework (SSLF-GR), and then simultaneously optimize the parameters and the ranking scores of graph nodes.

Sr. No.	Title	Author	Description
1	Data-Driven Answer Selection in Community QA Systems	Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang	To present a novel scheme to rank answer candidates via pairwise comparisons. In particular, it consists of one offline learning component and one online search component
2	Disease inference from health-related questions via sparse deep learning	L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T. S. Chua	A novel deep learning scheme to infer the possible diseases given the questions of health seekers
3	A ranking approach on large-scale graph with multidimensional heterogeneous information	B. Gao, T. Liu, T. Wang, G. Li, and H. Li	propose an innovative and effective semi-supervised Page Rank (SSP) approach to parameterize the derived information within a unified semi-supervised learning framework (SSLF-GR), and then simultaneously optimize the parameters and the ranking scores of graph nodes.
4	Beyond text QA: Multimedia answer generation by harvesting Web information	L. Nie, M. Wang, Y. Gao, Z. Zha, and T. Chua	The scheme that is able to enrich textual answers in cQA with appropriate media data. Our scheme consists of three components: answer medium selection, query generation for multimedia search, and multimedia data selection and presentation.

III. PROPOSED SYSTEM

An entirely distinct style for answer option in cQA setups. In the offline learning aspect, instead of lengthy as well as labour-intensive comment, the tendency to mechanically build the favorable, neutral, and also unfavorable training examples within the designs of choice sets target-hunting by our data-driven monitorings. Preliminary gather a pool of solution prospects through discovering its comparable questions. A significant variety of historic QA pairs, as time takes place, are archived within the cQA data sources. Information applicants for that reason have enormous opportunities to straight obtain the solutions by looking from the databases, rather than the lengthy waiting. A concern which is having several kinds of responses, however offering most appropriate solution of that question. Exact same kind of question which is addressed by specialists those experts will certainly be suggesting to individual for more questions.

A. SYSTEM ARCHITECTURE



The general work of the system is according to the accompanying:

1. User submit question on system
2. Online answer finding and expert recommendation
3. First remove the stopwords.

4. By using LDA algorithm find the topic of user question
5. Find expert for the same topic.
6. Offline search for answers
7. Search database for similar type of questions previously answered in database.
8. Perform PLANE model to rank each questions matching with users question.
9. Find positive, negative and neutral answers of those questions then eliminate negative question answers pair
10. Perform sentence level clustering with positive answers of matched questions and describe answer set of user's question .
11. Combined result of online and offline together to show result to user and recommend experts to user

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

IV. MATHEMATICAL MODEL

System – {I,P,O}
Input – I
Question – Q
Q = {Q1, Q2,…….Qn}
User – U
User = {U1, U2,…..Un}
Process - P
Step1 : -
User submit question Online
U → {Q1}
Step 2 :-
Remove stopwords from submitted question. Using LDA algorithm find the topic
Step 3 : -
Search related topic answers of that question
Step 4:-
Offline searching of an answer from database by using PLANE model to rank matching answers of the given question.
Step 5 :-
Find the positive, negative, neutral answers, for positive answers use sentence level clustering
Output o
Combined offline and online result
Result show to user
Recommend experts to user.

V. CONCLUSION AND FUTURE SCOPE

To provide an unique system for solution option in cQA setups. It consists of an offline learning and also an on the internet search element. In the offline learning element, rather than lengthy and also labor-intensive comment, An instantly build the favorable, neutral, unfavorable training examples the forms of choice sets directed by our data-driven monitoring. Then suggest a durable pairwise learning to rank model to integrate these 3 kinds of training examples. On the internet search part, for a provided question, initially gather a pool of solution prospects through locating its comparable questions. Then use the offline learned model to rank the solution prospects through using pairwise contrast. Actually to carried out comprehensive experiments to validate the efficiency of model on one basic cQA dataset and also one upright cQA datasets. In the end following factors: 1) The model could accomplish better efficiency compared to a number of cutting edge solution option standards; 2) The model is non-sensitive to its specifications; 3) The model is durable to the sounds triggered by increasing the size of the variety of returned comparable questions; 4) the pairwise learning to rank models including suggested PLANE are extremely conscious the mistake training examples. Past the standard pairwise learning to rank models, The model has the ability to integrate the neutral training examples as well as select the discriminative functions. It, nevertheless, likewise has the fundamental drawbacks of the pairwise learning to rank family members, such as noise-sensitive, massive choice sets, and also loss of info regarding the finer granularity in the significance judgment. In the future, to prepare to address such drawbacks in the area of cQA.

VI. ACKNOWLEDGMENT

With tremendous joy, distributing this paper as a piece of the educational modules of M.E. PC Engineering. It gives us pleased benefit to finish this printed material under the profitable direction of Principal for giving all offices and help for smooth advance of printed material. I want to thank all the Staff Members of Computer Engineering Department, Management, loved ones, Who have specifically or by implication guided and helped us for the planning of this paper and gives us a bolster ideal from the stage the thought was considered.

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, VOL. 29, NO. 6, JUNE 2017

[2] M. Ali, M. Li, W. Ding, and H. Jiang, Modern Advances in Intelligent Systems and Tools, vol. 431. Berlin, Germany: Springer, 2012.

[3] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T. S. Chua, "Disease inference from health-related questions via sparse deep learning," IEEE Trans. Knowl. Data Eng., vol. 27, no. 8, pp. 2107–2119, Aug. 2015.

[4] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor, "Learning from the past: Answering new questions with past answers," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 759–768.

[5] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in Proc. Int. Conf. Web Search Data Mining, 2008, pp. 183–194.

[6] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, "A framework to predict the quality of answers with non-textual features," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2006,pp. 228–235.

[7] Z. Ji and B. Wang, "Learning to rank for question routing in community question answering," in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage., 2013, pp. 2363–2368.

[8] T. C. Zhou, M. R. Lyu, and I. King, "A classification-based approach to question routing in community question answering," in Proc. 21st Int. Conf. World Wide Web, 2012, pp. 783–790.

[9] L. Yang, et al., "CQArank: Jointly model topics and expertise in community question answering," in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage., 2013, pp. 99–108.

[10] B. Li and I. King, "Routing questions to appropriate answerers in community question answering services," in Proc. 19th ACM Int. Conf. Inf. Knowl. Manage., 2010, pp. 1585–1588.

[11] K. Wang, Z. Ming, and T.-S. Chua, "A syntactic tree matching approach to finding similar questions in community-based QA services," in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2009, pp. 187–194.

[12] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in Proc. 31st Annu. Int. ACMSIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 483–490.

[13] M. J. Blooma, A. Y. K. Chua, and D. H.-L. Goh, "A predictive framework for retrieving the best answer," in Proc. ACM Symp. Appl. Comput., 2008, pp. 1107–1111.

[14] L. Nie, M. Wang, Y. Gao, Z. Zha, and T. Chua, "Beyond text QA: Multimedia answer generation by harvesting Web information," IEEE Trans. Multimedia, vol. 15, no. 2, pp. 426–441, Feb. 2013.

[15] Q. H. Tran, V. D. Tran, T. T. Vu, M. L. Nguyen, and S. B. Pham, "JAIST: Combining multiple features for answer selection in community question answering," in Proc. 9th Int. Workshop Semantic Eval., 2015, pp. 215–219.

[16] W. Wei, et al., "Exploring heterogeneous features for query focused summarization of categorized community answers," Inf. Sci., vol. 330, pp. 403–423, 2016.

[17] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton, "Quantitative evaluation of passage retrieval algorithms for question answering," in Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2003, pp. 41–47.

[18] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua, "Question answering passage retrieval using dependency relations", in Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2005, pp. 400–407.

[19] R. Sun, H. Cui, K. Li, M.-Y. Kan, and T.-S. Chua, "Dependency relation matching for answer selection," in Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2005, pp. 651–652.

[20] M. Surdeanu, M. Ciaramita, and H. Zaragoza, "Learning to rank answers on large online QA collections", in Proc. 46th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol., 2008, pp. 719–727.

[21] A. Agarwal, et al., "Learning to rank for robust question answering," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 833–842.

[22] D. Savenkov, "Ranking answers and Web passages for non-factoid question answering: Emory university at TREC LiveQA," in Proc. 24th Text REtrieval Conf., 2015.

[23] X. Li, Y. Ye, and M. K. Ng, "MultiVCRank with applications to image retrieval," IEEE Trans. Image Process., vol. 25, no. 3, pp. 1396–1409, Mar. 2016.

[24] M. K. Ng, X. Li, and Y. Ye, "MultiRank: Co-ranking for objects and relations in multi-relational data," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 1217–1225.

[25] D. H. Dalip, M. A. Gonc¸alves, M. Cristo, and P. Calado, "Exploiting user feedback to learn to rank answers in QA forums: A case study with stack overflow," in Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 543–552.

[26] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2010, pp. 411–418.

[27] J. Bian, Y. Liu, E. Agichtein, and H. Zha, "Finding the right facts in the crowd: Factoid question answering over social media," in Proc. 17th Int. Conf. World Wide Web, 2008, pp. 467–476.

[28] F. Hieber and S. Riezler, "Improved answer ranking in social question-answering portals, " in Proc. 3rd Int. Workshop Search Mining User-Generated Contents, 2011, pp. 19–26.

[29] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon,"Adapting ranking SVM to document retrieval," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2006,pp. 186–193.

[30] X. Li, G. Cong, X.-L. Li, T.-A. N. Pham, and S. Krishnaswamy, "Rank-GeoFM: A ranking based geographical factorization method for point of interest recommendation," in Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2015, pp. 433–442.

[31] J. Xu and H. Li, "AdaRank: A boosting algorithm for information retrieval," in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 391–398.

[32] X. Li, M. K. Ng, and Y. Ye, "MultiComm: Finding community structure in multi-dimensional networks," IEEE Trans. Knowl. Data Eng., vol. 26, no. 4, pp. 929–941, Apr. 2014.

[33] W. Wei, G. Cong, C. Miao, F. Zhu, and G. Li, "Learning to find topic experts in Twitter via different relations," IEEE Trans. Knowl. Data Eng., vol. 28, no. 7, pp. 1764–1778, Jul. 2016.

[34] W. Wei, B. Gao, T. Liu, T. Wang, G. Li, and H. Li, "A ranking approach on large-scale graph with multidimensional heterogeneous information," IEEE Trans. Cybern., vol. 46, no. 4, pp. 930– 944, Apr. 2016.

[35] X.-J. Wang, X. Tu, D. Feng, and L. Zhang, "Ranking community answers by modeling question-answer relationships via analogical reasoning," in Proc. 32nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2009, pp. 179–186.

[36] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in Proc. 16th ACM Conf. Conf. Inf. Knowl. Manage., 2007, pp. 919–922.

[37] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 221–230.

[38] V. R. Carvalho, J. L. Elsas, W. W. Cohen, and J. G. Carbonell, “Suppressing outliers in pairwise preference ranking,” in Proc. 17th ACM Conf. Inf. Knowl. Manage., 2008, pp. 1487–1488.

[39] Z. Zheng, K. Chen, G. Sun, and H. Zha, “A regression framework for learning ranking functions using relative relevance judgments,” in Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2007, pp. 287–294.

[40] X. Wei, H. Huang, C. Lin, X. Xin, X. Mao, and S. Wang, “Re-ranking voting-based answers by discarding user behavior biases,” in Proc. 24th Int. Conf. Artif. Intell., 2015, pp. 2380–2386.

[41] T. Joachims, “Optimizing search engines using clickthrough data,” in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 133–142.

[42] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, “Accurately interpreting clickthrough data as implicit feedback,” in Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2005, pp. 154–161.

[43] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, pp. 27:1–27:27, 2011.

[44] L. Nie, Y. Zhao, X. Wang, J. Shen, and T. Chua, “Learning to recommend descriptive tags for questions in social forums,” ACM Trans. Inf. Syst., vol. 32, no. 1, 2014, Art. no. 5.

[45] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in Proc. 31st Int. Conf. Mach. Learn., 2014,pp. 1188–1196.

[46] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.