INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 2)
Available online at: www.ijariit.com

An Efficient Community Question Answer (CQA) using the Metadata Rating

G. Divyatharshini
divyaganesh206@gmail.com
Tejaa Shakthi Institute of Technology for Women, Coimbatore, Tamil Nadu

P. Gokila Devi
gokiladevip333@gmail.com
Tejaa Shakthi Institute of Technology for Women, Coimbatore, Tamil Nadu

S. Srimathi
srimathi257@gmail.com
Tejaa Shakthi Institute of Technology for Women, Coimbatore, Tamil Nadu

A. M. SenthilKumar
senthil1185@gmail.com
Tejaa Shakthi Institute of Technology for Women, Coimbatore, Tamil Nadu

R. Sureshkumar
sweetsuresh17@gmail.com
Tejaa Shakthi Institute of Technology for Women, Coimbatore, Tamil Nadu

M. S. Vijaykumar
nklvijaykumar@gmail.com
Tejaa Shakthi Institute of Technology for Women, Coimbatore, Tamil Nadu

ABSTRACT

Community Question Answering (CQA) is the question answering process. Question Answering (QA) is the field of information retrieval and Natural language Processing (NLP). In this paper, we mainly focus on lexical gap and word embedding problem. Here, the CQA aims to find the existing question which is equivalent to queried question. The other existing system of this paper is two novel categories powered models. That is basic category powered model called as MB-NET and enhanced category powered model called as ME-NET. They are used to learn about the lexical gap and word embedding problem. The proposed system of this paper is to extend the metadata information easily using user ratings, like signals and Poll and Survey signals, into the learning process to obtain more powerful word representations.

Keywords: *Natural Language Processing (NLP), Information Retrieval, Community Question Answering (CQA) and Question Retrieval.*

1. INTRODUCTION

In a few years ago, the user-generated queries has become more important information resource on the web. That contains the Frequently Asked Question (FAQ) and Community Question Answering (CQA) such as Yahoo! Answers, Quora, Stack Overflow etc. The web contains the list of answers in the form of metadata. The metadata contains the category of the user's question and awards the best answers to the user. The CQA Archives valuable and various tasks like question-answering and knowledge mining. It is used to find the similar question or answer the queried questions. The best answer to the similar question is retrieved by using the knowledge of the CQA. The traditional retrieval ad hoc information retrieval, QA has several advantages firstly; it uses the natural language instead of the keyword as a query. Second, it has several possible answers instead of a ranked list and it is hard to find the required answers.

The lexical gap problem in CQA includes the previous work as it is divided into two types. The first is a traditional problem it has the question-answer pair to relate and learn the semantic relationship for the question -answers. It has the word to word and assuming the "parallel text" to obtain the similarity of the question-answer pair. Secondly, topic-based models, it is used to learn the topic based question-answer pair to get the lexical gap problem. It has the question and answer pair retrieves the same topic based model. It uses the skip-gram model, DEEPMATCH architecture [7], convolution neural network architecture (CNTNA) [8], to learn the word embedding and get the semantic information among the natural language questions. When the user asks the question in CQA then the user requires the category of the question and it is retrieved by using the predefined categories [4],[5],[9]. Here it has the

two models: one is basic category powered model and other is enhanced the category powered model. We build the regularization function to derive the metadata of category information along with the help of skip gram model using the training set. Once the word is embedded in continuous space then the question is viewed by using Bag-of-Embedded-Word (BoEW). The fixed length vector is aggregated by using the Fisher Kernel (FK) for the Variable –cardinality of BoEW.

2. RELATED WORKS

I. QUESTION RETRIEVED IN CQA:

In CQA researchers focus on metadata in CQA[1],[2],[3] to improve the performance o the traditional language model of QA. It has five groups of works. The first one is categories of questions. The framework integrates the category-specific term weight into the exiting VSM and BM25 retrieval model for question retrieval [10]. The combination of Global relevance and local relevance is used for question retrieval and VSM+WTLM shows the superior performance than other. CLM is used for viewing the category-specific term. The second one is question-answer pairs to learn the various translation models to bride the lexical gap problem. The word-based translation language model is used to learn the semantic similarity between the answers of the existing question and knows about the lexical gap by allowing the translation probability of similar questions. The third one is topic based modeling techniques to retrieving the best question and answer pairs. The fourth one is syntactic information for question retrieval. The fifth one is deep learning for question retrieval.

II. WORD EMBEDDING LEARNING:

The representation of the word is the continuous vector and it is attracted increasingly in the form of Natural Language Processing (NLP). The neural network language model (NNLM) is addressed in [15]. The efficient neural network word representation contains the continuous skip gram mode and continuous bag-of-word model (CBOC).

3. EXISTING SYSTEM

In this section, the proposed framework contains two steps. They are word embedding and fisher vector generation step. In the word embedding, it has the various NLP tasks. The skip-gram model [6] and continuous bag-of-words model (CBOW) [6] for learning word embedding it will free the memory. In the skip-gram model, it generates the training data. The softmax function is used:

$$P\ (w_{k+j}\ |w_k;\theta) = \exp\ (e^{T}_{w(k+j)}\ e_{wk})/(\sum^{N}_{w=1}\ \exp(e^{T}_{w}e_{wk}))$$

3.1 Basic Category Powered Model (MB-NET):

The metadata powered model is the next step of the skip gram model. It contains the information about the metadata and hierarchy of the category[4],[10]. It is used to find the similarity between the existing question and the newly queried question of the users.

Here, the regularization function is used:

$$S_b(w_k,w_i,c_k)= \begin{cases} 1\ if\ ck = ci \\ 0\ \ otherwise \end{cases}$$

$$E_b\ = \sum_{k=1}^{N} \sum_{i=1}^{N} S_b(w_k,w_i,c_k)d(w_k,w_i)$$

3.2 Enhanced Category Powered Model (ME-NET):

This is used to overcome the problem of the basic category powered model. It is used to exist the similarity of leaf categories under the same category[17].

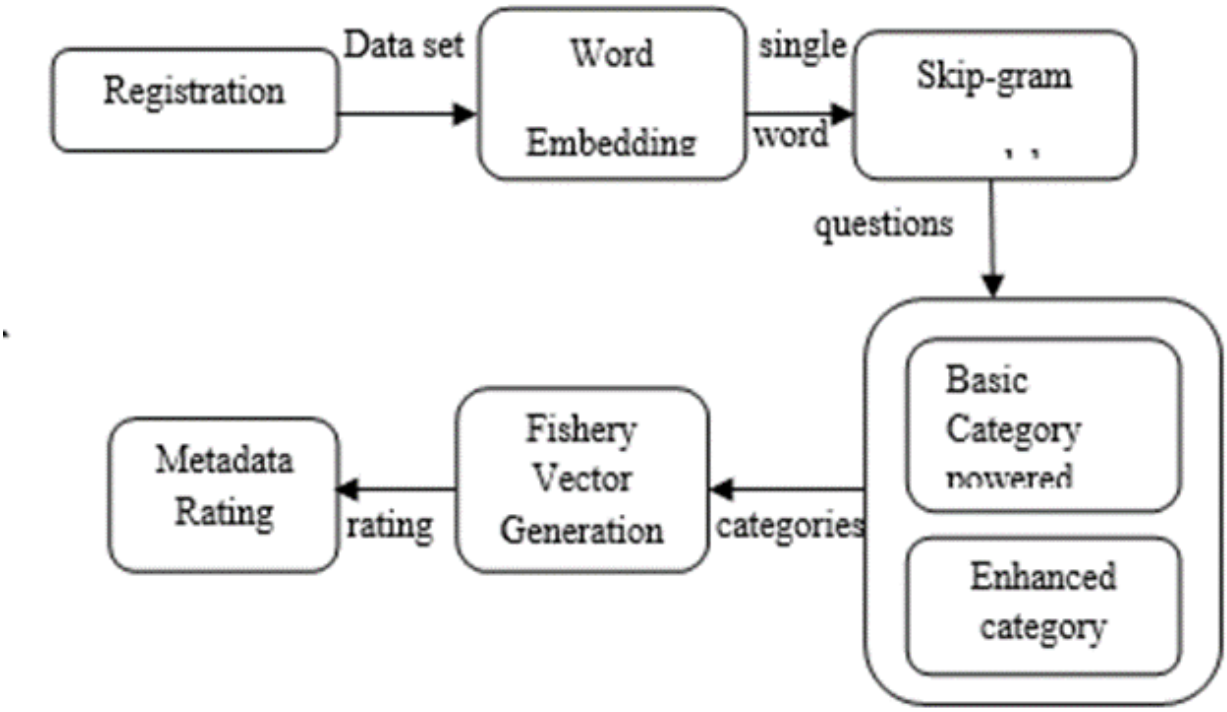
$$S_e(w_k,w_i,c_k) = 1/A\ \{\gamma s_b\ (w_k,w_i,c_k) + \sum_{ci\in Related(ck)} R(c_i\rightarrow c_k)s_b(w_k,w_i,c_i)\}$$

3.3 FISHERY VECTOR GENERATION:

Here, the word embedding is learned and the question can be a representation is by using the BoEW. The semantic level similarity between queried questions and the existing questions represented by BoEW can be captured more accurately by BoWs method. It uses the Gaussian mixture model (GMM). In GMM, the λ is implemented by Maximum Likelihood (ML) using the Expectation Maximization (EM). The final derivation is

$$G^{qi}\sigma_{,k} = \frac{1}{Ni\sqrt{2\theta i}} \sum_{j=1}^{Ni} \gamma_k\ (k)\ [((ewij\text{-}\mu_k)^2/\sigma_k^2)\ \text{-}1]$$

4. PROPOSED SYSTEM



In the proposed system we have a metadata information can easily retrieve the information based on the user rating like poll and survey signals.

In the proposed system, it has the set of questions as the input. And it is used to addresses the problem of the lexical gap and word embedding. Here, we have the word embedding as two processes they skip gram model, basic category powered model and enhanced powered model. The basic model is used to categories the given user query and the enhanced model is used to overcome the problem of the basic category model.

That uses the regularization function, to solve the category problems and it checks the existing question and newly queried question. It is the process of getting the question as input and checks whether the question is already existed and provide the best-ranked answer for the users.

It is the process of the getting the training set by using the skip-gram model. The skip gram model is used to retrieve the answer easily and saves the memory. After, skip gram the MP-NET is used to retrieve the best category of the queried questions. Then the fishery generation is used it is generated by using GMM model. And the proposed system to generate the rating process. The Metadata information is rated by using the users rating process. So, the more rated answer is viewed to the queried questions.

The question is given as input to the proposed system and making the skip gram model to get the training set of date. After, that the category is retrieved by using the existing process and the proposed system is used to retrieve the best Question-Answer pair for the user's questions.

5. DATASETS

We collect the data set in the Yahoo! Answers and that is called as the retrieval data or training data. The retrieved question contains the title, description, and category in the metadata. In order to create the dataset, we collect the extra question and that all are posted more recently. Hence, BM25 is used for ranking the queried question and it is stored in the metadata information. The data set is divided into two sets they are validate set and test set. The validate set is used for tuning parameter the process and the test set is used to evaluating the ranked model how the relevant candidate is a contrast to irrelevant data.

In order to evaluate the performance of the different models. Here, they use the different measures are Mean Average Position(MAP), Mean Reciprocal Rank(MRR), R-Precision (R-Prec) and Precision at N(P@N).[4]

These are used to measure the question retrieval in CQA.

MAP Measures the mean average precision for the queried question.

$$\text{MAP} = \frac{\sum q \epsilon Q \; Avg\; P(q)}{|Q|}$$

$$\text{Avg P(q)} = \frac{1}{Nmq} \sum_{j=1}^{|Mq|} \frac{NMq,i}{j} 1(M_{q,j})$$

MRR is used to evaluate the list of possible response for the set of queries.

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{ranki}$$

ADVANTAGE OF PROPOSED SYSTEM

- The metadata information can be retrieved easily using rating.
- The information of the user can be obtained accurately and gives a better performance.
- It is very fast to access the information for the user.

DISADVANTAGE OF EXISTING SYSTEM

- It does not retrieve the information based on existing metadata.

6. CONCLUSION

In this paper, we proposed to learn about the continuous vector representations for question retrieval in CQA. For that, we have the addresses the lexical gap problem and word embedding problem. After that we have the novel based models: one is basic category powered model (MB-NET) and enhanced category powered model (ME-NET). Once the word is embedded into continuous space then that will be treated as BoEW.

The metadata information is rated to retrieve the best pair of question an answer. The metadata is rated using the best-ranking process. It can be easily extended to incorporate other metadata information, such as user ratings, like signals and poll and survey signals.

7. REFERENCES

[1] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in the large question and answer archives," in Proceedings of the CIKM, 2005, pp. 84–90.
[2] X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in Proceedings of the SIGIR, 2008, pp. 475–482.
[3] J.-T. Lee, S.-B. Kim, Y.-I. Song, and H.-C. Rim, "Bridging lexical gaps between queries and questions on large online q&a collections with compact translation models," in Proceedings of the EMNLP, 2008, pp. 410–418.
[4] X. Cao, G. Cong, B. Cui, and C. S. Jensen, "A generalized framework of exploring category information for question retrieval in community question answer archives," in Proceedings of the WWW, 2010, pp. 201–210.
[5] L. Cai, G. Zhou, K. Liu, and J. Zhao, "Learning the latent topics for question retrieval in community qa," in Proceedings of the IJCNLP, 2011, pp. 273–281.
[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the NIPS, 2013, pp. 3111–3119.
[7] Z. Lu and H. Li, "A deep architecture for matching short texts," in Proceedings of the NIPS, 2013, pp. 1367–1375.
[8] X. Qiu and X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in Proceedings of the IJCAI, 2015, pp. 1305–1311.
[9] X. Cao, G. Cong, B. Cui, C. S. Jensen, and C. Zhang, "The use of categorization information in language models for question retrieval," in Proceedings of the CIKM, 2009, pp. 265–274.
[10] Z. Ming, T. Chua, and G. Cong, "Exploring domain-specific term weight in archived question search," in Proceedings of the CIKM, 2010, pp. 1605–1608.
[11] G. Zhou, Y. Chen, D. Zeng, and J. Zhao, "Towards faster and better retrieval models for question search," in Proceedings of the CIKM, 2013, pp. 2139–2148.
[12] S. Clinchant and F. Perronnin, "Aggregating continuous word embeddings for information retrieval," in Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, 2013, pp. 100–109.
[13] J. Sanchez, F. Perronnin, T. Mensink, and J. J. Verbeek, "Image classification with the fisher vector: Theory and practice." International Journal of Computer Vision, pp. 222–245, 2013.
[14] G. Zhou, T. He, J. Zhao, and P. Hu, "Learning continuous word embedding with metadata for question retrieval in community question answering," in Proceedings of the ACL, 2015, pp. 250–259.
[15] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," J. Mach. Learn. Res., vol. 3, 2003.
[16] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa, "Natural language processing (almost) from scratch," Journal of Machine Learning Research, vol. 12, pp. 2493–2537, 2011.
[17] B. Li, I. King, and M. R. Lyu, "Question routing in community question answering: putting category in its place," in Proceedings of the CIKM, 2011, pp. 2041–2044.
[18] G. Zhou, S. Lai, K. Liu, and J. Zhao, "Topic-sensitive probabilistic model for expert finding in question answer communities," in Proceedings of the CIKM, 2012, pp. 1662–1666.
[19] G. Zhou, Y. Chen, and D. Zeng, "Group non-negative matrix factorization with natural categories for question retrieval in community question answer archives," in Proceedings of COLING 2014, 2014, pp. 89–98.
[20] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in Proceedings of the SIGIR, 2001, pp. 334–342.

[21] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu, "Statistical machine translation for query expansion in answer retrieval," in Proceedings of the ACL, 2007, pp. 464–471.
[22] G. Zhou, K. Liu, and J. Zhao, "Exploiting bilingual translation for question retrieval in community-based question answering," in Proceedings of COLING 2012, 2012, pp. 3153–3170.
[23] G. Zhou, F. Liu, Y. Liu, S. He, and J. Zhao, "Statistical machine translation improves question retrieval in community question answering via matrix factorization," in Proceedings of the ACL, 2013, pp. 852–861.
[24] W. Zhang, Z. Ming, T. Liu, and T. Chua, "Capturing the semantics of key phrases using multiple languages for question retrieval," IEEE Trans. Knowl. Data Eng., vol. 28, no. 4, pp. 888–900, 2016.
[25] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the potential of q&a community by recommending answer providers," in Proceedings of the CIKM, 2008, pp. 921–930.

