Advances in Parallel Computing Technologies and Applications
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC210118

Auto Question Tagging for Health Care Using Machine Learning Technique

Kathiravan M [a,1], Irumporai A [b], Sreesubha S [b] and Madhurani M [b]
[a] *Hindustan Institute of Technology & Science, Padur, Chennai, 603 103, India*
[b] *Rajalakshmi Engineering College, Thandalam, Chennai, 602 105, India*

Abstract. Tagging is a machine learning technique that provides tags to the information that the user can easily identify the related information. Manual tagging is widely used for constructing question banks; but, this approach is time-consuming and it would create pathway to consistency issues. Semi-manual tagging which is time-consuming and people must be experts in that domain have the ability to identify the question and tag them it is not possible in real-time and high in cost. The proposed associate degree automatic tagging exploitation information processing that mechanically tags automatic question tagging. In this paper, step up with a keywords-based model to automatically tag questions with information units. With regard to multiple-choice questions, the proposed models use mechanisms to capture helpful information from keywords to boost tagging performance. Automatic tagging method using NLP automatically tag questions where users can get information regarding to his search which overcomes earlier methods. In our experiments, the result shows that the model is credible and outperforms various existing models.

Keywords. Machine Learning, Natural Language Processing, Clustering, Prediction, E-Learning.

1. Introduction

With science of innovation, a modern computerised world has become multisource environment where one can get any information from www. While many applications were focus on extracting information from web, no technique has found to help user for identifying relevent information. The increased population of today young community might prefer to get most of the study material from open web for their academic purpose. While the usage of web increases for academic studies, finding or organizing relevant information become a mandate for medical too. But, the issue is that relying on web sources for medical treatment is invalid. Therefore, there has been a well structured methods has come to power to help user to get appropriate information from medical expert. Though the existing methods are efficient to retrieve accurate information from web, the integrated question answering approach requires to be integrated so that user can post any question and get right medicine from experts through online. Likewise, the well organised question bank with answers necessary to get aware of their field and raise relevant

[1] Dr.M.Kathiravan, Department of CSE, Hindustan Institute of Technology & Science.
Email: kathirrec1983@gmail.com

question to expert. Therefore, a well organized computerized adaptive testing method has proposed for providing right information to the user according to the individual user interest. This method facilitates a collection of questions stored in the system where tagging help to organize the resources. Since the tag has potential to organize knowledge units with required answer, the questions can be easily associated with relevant one. Thus, gives sophisticated CAT functions. Tagging is a machine learning technique which provides tags to the information that user can easily identify the related information user searching for. But there is flaw in that method, most technology uses manual tagging and semi-manual tagging which is time consuming and quiet difficult in real time and high in cost. An automatic tagging method by NLP automatically tag question where user can get information regarding to his search which overcome earlier methods.

2. Literature Survey

In variant from the article stated [1], dynamic model has been employed and secured best result over this existing model. The knowledge has been extracted from millions of questions from the stack overflow of Q&A site. The work was achieved by discriminative approach where there was no trial data set enforced. The work from [2], demonstrated by utilizing local mining and global study technique. The main issue in this approach is that it leads to data loss and low precision rate. The statistical model that was used [3] were trained on huge data particles graph to predict unknown labels. The observation on this existing article is that no text information is useful for developing knowledge in web. Similar work was carried out [4], by automatic labelling of textual data with knowledge base. The weakness is that the search queries were not found accurately as well as it generates irrelevant answers mostly. The idea from [5–7] tightly bonded and strengthened mutually. It does not carry the data dynamically since the idea is static and the approach is rudimentary and incompetent. The work [8,9] is mainly concentrated for schools where pretrained question were posted to school students. It lacks with automatic tagging semantically in the World Wide Web. The authors proposed [10], a system to focus only on context dependent matters so that it fails to adapt complicated Quiz and related sets. The work that was carried in our approach was novel and supports quiz related question and answering. The dropout approach [11] addresses on random bead piece on neural network platform during trial. Though this method was improving the performance on supervised learning and data modelling it generates huge noise which affects the overall performance. The work that has been reported from [12–16], is similar and our method was achieved very efficient for convolution implementation. The work in our approach reduces the overfitting issue that was mainly degrading the performance of the existing work.

3. Proposed Method

In this system, an automatic tagging of questions is performed by NLP-based machine learning (Natural language processing) technique. Machine Learning (ML) is applications of artificial intelligence (AI) that provide systems the ability to automatically learn. NLP analyzes to understand the human language in a smart and useful way. The Medical

datasets are collected from the Medinet library. The data are stored in the CSV file formats for later use. CSV file contains the questions and answers from the medical health domain and then the automatic tagging takes place of the question and Answer than the user search the query in AWS(Amazon web service) and the query is forwarded to the Medinet library to extract the user queried related answer. Medinet library contains 700 medical domain files which are stored in tc2011 API if the requested file not found in Medinet it redirects to PDF box it will provide pdf files(contains information about domain or diseases) related to the question. The pdf files are extracted by using Lucene indexing which provide fast retrieval of files if the answer is not related to query or not clear. The question will be forwarded to the Expert (doctor) who can clear the doubts and replay for the query. Finally providing the quiz for the student who wants to know their knowledge in the medical field and providing them domain score and overall score and feedback for the student. The proposed methodology is applied in order to answer the medical related queries in order to reduce the confusion in medical field. It also avoids the user from getting distracted what they originally need. Thus, encourages people and students to gain valid information regarding the medical field. The system can be used in websites. It provides standard and consistent results. There are various modules that perform different task in our proposed methodologies.

3.1. Admin Pre-processing

First module Admin has to register first and then provide admin details (Name and password) valid admin enters the page. Admin is responsible for the whole operation admin work includes cleaning, adding the CSV files, analyzing resources, NLP (Natural language processing) and cleaning NLP.

3.2. Auto tagging Questions

This module helps user to register, authenticate and access relevant information from choosing the list of domains available in the database. The answer can be supplied by the expert for the question posted by the user if the relevant answer is not available from back end. If the user needs more clarification on this result can interact with expert and get details as much as needed.

3.3. Expert's Answering Process

This is expert module which can be accessed by only authorized or registered domain expert. The primary purpose of this module is to provide accurate answer for all the user queries while communication taken place between user and expert. The expert can interact or sophisticate the user by proving truly valid information over the chat or message. The expert answer is prepared after thorough analysis of user queries and involved two level authentications from the expert in order to validate the expert recommendation.

3.4. Student Assessment

In this module student registers initially, after that login page would forward to the quiz page the students were asked to write quiz in medical domain. Questions from medical

domain would provide to student and they can answer it. Finally, the overall score of the test, domain score, and the feedback for the test would be provided, and then the student could improve knowledge in that domain.

4. Experimental Evaluation

Though some studies were focussed on this issue for some business agenda, question bank with knowledge tags present at the moment. In this circumstance, a well-executable method was applied to construct MCQ based question bank in our implementation. The question was crawled from a famous website such as koolearn3 and Tiku4, which returns expected results through search engine. In our initial process the questions that are similar to others are removed. The question that is collected from web is based on English subject on elementary and higher secondary school. The prototype was constructed with using English syllabus and curriculum in which knowledge map is coordinated by ontology based hierarchical knowledge system. To label the question tag leaf information is used. The trial experiment shows the consistency from the training sets. The F-1 score obtained from training is 0.8484 which shows that the pattern that has been extracted from this approach is not hard for the proposed language. The Kappa and F1-score are 0.4753 and 0.8996, respectively. The notable observation here is that a Kappa is iterative so we achieved 0.8197 in our final experimentation. This result obtained in our experiment is best compared to other metrices.

Table 1. Evaluation results of medical terminology assignment in terms of csv and expert answers.

Approach or Metric	S@1	S@2	S@3	S@4	P@1	P@2	P@3	P@4
CSV	72.0%	84.0%	91.0%	95.0%	72.0%	71%	69.7%	68.3%
EXPERT	83.0%	92.0%	98.0%	100.0%	83.0%	81.5%	80.3%	8.8%

Table 2. Comparative illustration of the representative question samples with csv and expert answers.

QA pairs	CSV ANSWERS	EXPERT ANSWERS
Do I suffer by hair colouring while pregnancy?	hair structure, dyed hair, feeling safe, patient currently pregnant, first trimester pregnancy...	Hair structure, coal tar allergy, hair disorder of endocrine system.
If I get an infection caused by gum disease, can that be transferred to my fetus?	Gingival disease, inflammation, periodontal disease...	Prematurity of fetus, gingival disease, periodontal disease low birth weight infant...

5. Conclusion and Future Work

This work achieves superior performance in indentifying relevant medical information through online. The advancement in search and querying methods checked out parametric conditions and necessities of user's needs accurately for today and tomorrow. Maintaining and keeping up accuracy to a standard is always tougher and troublesome with time. Some of the challenges can be anticipated, such as advances in algorithmic con-

versions that are making it easier to provide accurate search results from the database. Here predictive algorithm is being utilized to work around the basic shortcomings in user search results. As the confirm mechanism for searching our view could be suitable and accurate. Similarly, we have also developed a system that does not allow any user from getting diverted from the search they wanted to perform.

References

[1] Al-Hmouz A, Shen J, Al-Hmouz R, Yan J. Modeling and simulation of an adaptive neuro-fuzzy inference system (ANFIS) for mobile learning. IEEE Transactions on Learning Technologies. 2011 Dec 13;5(3):226-37.

[2] Brusilovsky P. KnowledgeTree: A distributed architecture for adaptive e-learning. InProceedings of the 13th international World Wide Web conference on Alternate track papers & posters 2004 May 19 (pp. 104-113).

[3] Mendicino M, Razzaq L, Heffernan NT. A comparison of traditional homework to computer-supported homework. Journal of Research on Technology in Education. 2009 Mar 1;41(3):331-59.

[4] Forbey JD, Ben-Porath YS. Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. Psychological assessment. 2007 Mar;19(1):14-24.

[5] Nickel M, Murphy K, Tresp V, Gabrilovich E. A review of relational machine learning for knowledge graphs. Proceedings of the IEEE. 2015 Dec 17;104(1):11-33.

[6] Luo Y. Environmental cost control of coal industry based on cloud computing and machine learning. Arabian Journal of Geosciences. 2021 Jun;14(12):1-6.

[7] Xu X. Machine learning-based prediction of urban soil environment and corpus translation teaching. Arabian Journal of Geosciences. 2021 Jun;14(11):1-5.

[8] Kim HL, Passant A, Breslin JG, Scerri S, Decker S. Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces. In2008 IEEE International Conference on Semantic Computing 2008 Aug 4 (pp. 315-322). IEEE.

[9] Brut M, Sedes F, Jucan T, Grigoras R, Charvillat V. An ontology-based modeling approach for developing a competencies-oriented collective intelligence. InIFIP World Computer Congress, TC 3 2008 Sep 7 (pp. 219-222). Springer, Boston, MA.

[10] Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering. 2013 Mar 7;26(8):1819-37.

[11] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM. 1975 Nov 1;18(11):613-20.

[12] Saha AK, Saha RK, Schneider KA. A discriminative model approach for suggesting tags automatically for stack overflow questions. In2013 10th Working Conference on Mining Software Repositories (MSR) 2013 May 18 (pp. 73-76). IEEE.

[13] Ambika M, Raghuraman G, SaiRamesh L. Enhanced decision support system to predict and prevent hypertension using computational intelligence techniques. Soft Computing. 2020 Feb 14:1-2.

[14] Saranya MS, Selvi M, Ganapathy S, Muthurajkumar S, Ramesh LS, Kannan A. Intelligent medical data storage system using machine learning approach. In2016 Eighth International Conference on Advanced Computing (ICoAC) 2017 Jan 19 (pp. 191-195). IEEE.

[15] Ambika M, Raghuraman G, SaiRamesh L, Ayyasamy A. Intelligence–based decision support system for diagnosing the incidence of hypertensive type. Journal of Intelligent & Fuzzy Systems. 2020 Jan 1;38(2):1811-25.

[16] Wang L. Urban land ecological evaluation and English translation model optimization based on machine learning. Arabian Journal of Geosciences. 2021 Jun;14(11):1-6.