A Cluster-based Approach for Finding Domain wise Experts in Community Question Answering System

R Menaha¹, VE Jayanthi², N Krishnaraj³, N Praveen sundra kumar⁴
1, 3, 4 Asst.Professor, IT department, MCET, Pollachi.
2 Professor, ECE Department, PSNACET, Dindigul
1.rmenahasenthil@gmail.com ³ krishnaraj@drmcet.ac.in ⁴ npsk.2020@drmcet.ac.in

Abstract. Community Question Answering (CQA) systems is an emerging web-based information service. CQA enables web users to get precise answers to questions from experts of the specific domain. CQA is used in wide areas such as biomedicine, information technology, tourism, etc. This paper focused on finding experts in community question answering system using unsupervised machine learning technique. Our Proposed system consists of three phases namely i) Clustering tags ii) Determining Experts for the unanswered questions iii) Finding experts for the given Question. By doing the tag analysis process, identified similar tags in a particular domain, and formed a cluster. By doing question analysis, we found the unanswered questions in each domain and identified the experts. For the given question, suggest the expert by doing pattern matching technique. The results section proves that our system predicts the experts for the given question with good accuracy.

1. Introduction
Question answering is an information retrieval system that aims to provide direct answers for the query from the structured database of knowledge or information. Generally, two types of question answering (QA) systems exist on the web. i) Closed-domain ii) Open domain. The closed domain QA system deals with questions beneath a specific domain. The open-domain QA system deals with questions in a wide area. An open-domain QA system is developed in Menaha et al. [11] using web snippets as a knowledge base. Another open domain voice-based QA system is developed in Maheshwari et al.[9] by exploiting NLP techniques and used Web-snippets as a knowledge source. In recent times, the QA system is developed for many real-time applications such as medical, e-learning, project management, community question answering (CQA), etc. In particular, CQA forums are gaining popularity online.

CQA websites help people to ask their queries instead of searching for an answer in a search engine. While searching a question in the search engine gives millions of results as web pages but the CQA website gives the direct and relevant answer to the posted queries. So the user’s time is saved for exploring the data. Nowadays, CQA websites like Yahoo Answers, Quora, StackOverflow, StackExchange, Wikianswers are used by many users to search and explore knowledge. For example, the users can ask questions in the information technology domain and the experts post their answers for the queries asked by other users.

Usually, in CQA websites, the user can ask their question and n number of experts will post an answer for that question based on their knowledge in that domain. The user who reads that answer can provide their vote in two forms namely upvote and downvote. For the most relevant answers, it is
upvoted, whereas the less relevant or irrelevant answer is downvoted by the user. Based on the voting, the most relevant answer can be known to the user easily.

In CQA websites, previously asked questions and answers are available in archives so that the user can find the answer sooner instead of waiting for other users to post the answers. These websites are helpful for the user to find the best answer for their question and find the experts in that domain. In few CQA websites, flags are used to identify whether the user is good or bad but it may cause deviation. For finding the best answer, the user should have more knowledge related to that domain. And for the finding experts, the user needs to analyze the answerer profile.

This work is mainly focusing on finding experts in the CQA website for the identified tag from the given question. Our main contributions in this work are summarized as follows:

- As an initial work, a tag analysis process is performed on CQA websites to identify the technical areas where the user asking questions and posting answers. For each such domain, similar tags are identified from the CQA website. These tags are grouped as a cluster. We have formed an n x m matrix to identify the experts for each tag where n denotes the user and m denotes the tag.
- In the Question Analysis process, we have collected the total number of answered and unanswered questions for each cluster.
- To suggest the expert for the given question, data preprocessing is done, and identified the keyword from the given question. By using a string matching algorithm, the extracted keyword and the tag in the cluster are compared from which the experts are recommended to the user.

This paper is organized as follows. Section 2 presents related work and different approaches followed in CQA. Section 3 gives the problem definition and solution overview. The section 4 describes our proposed methodology. Section 5 shows the results and experiments of the system. Section 6 concludes the proposed system.

2. Related Work

In the past decade, several works were made by using Community Question Answering sites. Some of the works include routing questions to the experts, finding question quality, determining good users or bad users, finding high-quality answers from archives, identifying best answers, discovering experts, and high-quality posts. Those works are presented in this section.

Baichuan Li et al. [4] presented a method to route questions to the experts of the domain based on their previous answering profile. They have developed a framework called question routing which consists of four phases namely performance profiling, expert estimation, availability estimation, and answerer ranking. Antoaneta Baltadzhieva et al. [3] presented an approach to find the question quality based on features such as tags, question title, question body length. They have used an algorithm called supervised latent dirichlet allocation for classification purposes.

Imrul Kayes et al.[8] classified the user as Good users or bad users based on the flags in the CQA website. Hapnes Toba et al.[7] offered an approach to find high-quality answers from the archives. Maria Soledad Pera et al. [10] presented a method to choose the top-ranked related answer by using QAR.

Wenhao Zheng et al. [14] provided a method to identify the best answer in the CQA forum by using heterogeneous sources. Dallia Elafy et al. [5] delivered a hybrid model to predict the best answers by using content and non-content features. Roy et al. [13] classified the answers based on their recent occurrences by considering time and post.

Fatemah Riahi et al. [6] presented a method for finding experts using a segmented topic model. Zhenlei Yan et al. [16] provided two models called tensor and topic models to rank the optimal answerer based on the optimized AUC (Area Under the ROC Curve). Yuan Yao et al [15] found the high-quality question and answer by considering the feedback based on voting.
Abdelghani Bouziane et al. [1] surveyed question answering systems. Similarly, Amit Mishra et al. [12] presented a survey on question answering systems with classifications. Annamoradnejad et al. [2] presented a method for predicting features from QA websites using BERT.

From the survey analysis, it is observed that there are no domain wise analysis and experts finding for each domain on the CQA website. Moreover finding experts for unanswered questions in CQA is not yet experimented with. To address the above-said limitations, we have proposed an unsupervised machine learning model to find the experts for the particular domain and found the experts for the unanswered questions.

3. Problem Definition

This section presents the problem definition and solution overview.

3.1 Problem Definition

Definition 1: Identify the domain \( d_1 \ldots d_n \) where more number of questions and answers are posted from the CQA website. Form the clusters \( c_1 \ldots c_n \) by grouping relevant tags in that domain.

Definition 2: For each cluster \( c_1 \ldots c_n \), Identify the experts by doing the expert analysis process.

Definition 3: Identify the expert for the given question by analyzing the cluster data.

3.2 Solution Overview

Initially, domains are identified from the CQA website, for each domain collected relevant tags and formed a cluster for each domain. For each cluster experts are identified by collecting details from the CQA website in the form of a matrix. Each cluster is implemented as a hash map structure i.e. key-value pairs where key meant tag and value meant expert. For the given question, the expert is recommended by analyzing cluster data by using the bag of words model. The novelty of our approach is recommending an expert for the given query based on domain.

4. Methodology

The proposed system is implemented as two phases of work. i) Clustering similar tags ii) Finding experts for the given question. The outline of our system implementation is illustrated in fig 1.

4.1 Clustering Similar Tags

Here, the clusters are formed based on three steps of the process: i) Domain Identification ii) Tag Analysis and iii) Expert Analysis. Each step is discussed in this section.

4.1.1. Domain Identification

The Stack Overflow CQA website is used for doing experiments. An analysis is made on the CQA website for finding the technical areas where the users raise more queries and answerer post their answers. The set of identified technical areas are called domains. The identified domain are Java, Web, Python, .Net, SQL, Django, R, Angular, Apple, Excel, OS, ASP.Net, Entity Framework, C, JavaScript, Git, Qt, Database, Android, Ruby, and others.

4.1.2. Tag Analysis

A tag is a keyword or label that categorizes our question with other similar questions. For each domain, we have collected similar tags from Stack Overflow. These tags are grouped as a cluster with the domain name. A cluster may consist of many tags. For example, the cluster Java consists of tags
Figure 1. Outline of the Proposed System

such as Java, JSF, JSF 2, Servlets, Swing, JPanel, JTable, JSP, Java 8, Java Stream, Jim, Garbage-Collection, JPA, JUnit, Spring, Spring-MVC, and Hibernate.

4.1.3. Expert Analysis
We have analyzed the user pages in Stack Overflow to identify the experts in each domain. This page contains the details of user name, Number of answers posted for a tag by that user, and their score for a tag. By using this data, we have formed a matrix (n x m), where n denotes the user and m denotes the tag which is used by the user. For every single user, the score for each tag is recorded in a matrix format. By using this matrix, we have identified the answerers in that tag and also the top scorer for that tag.

From this matrix, the experts for each tag in a cluster are found based on their score in the CQA site. An expert is a person who has the highest score in that particular tag for the domain. In a cluster, the tags are arranged in the form of key-value pairs, where the key is the tag and the value is the expert for that tag. A single tag can have more than one expert at the same time a person may be an expert in more than one domain.

4.2 Finding Experts
Usually, more number of answerers or experts are available in Stack Overflow, but still millions of queries are yet to be answered by the experts. As an example, in the java domain 267380 queries, in the web domain, 690892 queries are yet to be answered. To provide support, we have collected several unanswered questions in each domain from stack overflow.
Finding experts for the given question is implemented as four steps in our proposed system. The steps are i) Question Preprocessing ii) Determining Experts, which are discussed in section 4.2.1 and 4.2.2 respectively.

4.2.1 Question Pre-processing

Question Pre-processing is an important step in the analysis process. The pre-processing technique involves transforming raw data into an understandable format. Normally, real-world data is often incomplete, inconsistent and this may cause errors. Pre-processing is a method of resolving such errors or issues. In pre-processing, the question should be analyzed otherwise it may lead to an improper result. The quality of the question is first and foremost before running an analysis. The higher quality question gives a higher quality answer.

As a question pre-processing process, the given question is tokenized i.e the words are split as tokens. Then the stop words from the question such as what, who, which, where, then, is, are, uses, that, these, those, at, in, on, by, how, whom, if, such, when, this, etc., are removed. Therefore the important words in the given question are left to remain. Then the important keywords are used for mapping with domain in the cluster.

4.2.2 Determining Experts

For finding the expert, we have used the concept of a hash map. Hash map is the collection of key-value pairs. It maps keys to values. The hash map names remain the same as the cluster name. There are twenty-one hash maps are available in our system.

In a cluster, each tag is represented as key-value pair whereas key refers to the tag and value refers to the expert for that tag. A single cluster can contain more than one key-value pair. Once the keywords are obtained from the question pre-processing, these keywords are mapped with each key of the cluster. If the keyword matches the key of the cluster, it suggests the expert for that tag as a result. The following equation (1 & 2) expresses the keyword and tag matching.

\[
f(e) = \sum_{k=1}^{n} \sum_{c=1}^{m} \sum_{t=1}^{n1} f(t) \quad \ldots \quad (1)
\]

\[
f(t) = \begin{cases} 
1 & \text{If } k = t \\
0 & \text{else} 
\end{cases} \quad \ldots \quad (2)
\]

In Eqn. (1), the function \(f(e)\) is used for finding experts from each cluster where \(k\) refers to the keyword, \(c\) refers to the cluster, and \(t\) refers to the tag in each cluster. Here, \(n\) refers to the number of keywords, \(m\) refers to the number of clusters and \(n1\) refers to the number of tags in each cluster.

In each cluster, the tags are examined and compared with the keywords of the given question. The function \(f(t)\) compares the keyword with the tag. In Eqn. (2), \(k\) refers to keyword and \(t\) refers to tag. If there is a match, then it returns 1 otherwise return 0. If the \(f(t)\) value is 1 then the value of the tag is the corresponding expert.

5. Experiments and Results

We have used the Stack Overflow CQA website for experimental purposes. The system is implemented by using java language. Initially, domain analysis is done to find out the technical areas where more questions are thrown by users and posted by the experts. As a result, we have identified 21 clusters.
5.1 Data Collection

For each domain, we have gathered the relevant tags from Stack Overflow. For example, by giving the search query python in Stack Overflow, it returns the relevant tags such as Python, Python -2.7, Python 3.x, Numpy, List, Dictionary, Py.test, Pip, Pandas, DataFrame, etc. By doing this process, we have collected the relevant tags in each domain. Table 1 shows the relevant tags in each domain. In table 1, C1…C21 refers to the number of identified clusters. More than 200 user pages are explored from the user page in Stack Overflow. On each page, the user score is extracted for each tag. By doing this process, we have formed an n * m matrix, where n refers to tag and m refers to users or answerers. From the formulated matrix, the list of answerers for each tag is identified by using their score. The top answerers for each tag are also identified based on their score. Table 2 shows the list of answerer and the top answerer for each tag. In Table 2, A1, A2,…An refers to the answerers or experts.

Table 1. Tags in each Cluster

| Cluster | Domain | Relevant Tags |
|---------|--------|---------------|
| C1      | Java   | Java, Jsf, Jsf 2, Servlets, Swing, JPanel, JTable, Jsp, Java 8, Java Stream, JVM, Garbage-Collection, JPA, JUnit, Spring, Spring-MVC, Hibernate |
| C2      | Web    | CSS, Html, WordPress, Cart, Rest, Jax-rs, Html5, CSS3, Flexbox, CSS selector, WooCommerce, Products, PHP, Regex, google-chrome-extension, google-chrome, Maven |
| C3      | Python | Python, Python -2.7, Python 3.x, Numpy, List, Dictionary, Py.test, Pip, Pandas, DataFrame |
| C4      | .net   | .net, String, Generics, C#, Linq, Visual studio, WinForms, Vb.net |
| C5      | SQL    | MySQL, SQL-server, SQL, PostgreSQL, TSQL, SQL server 2008, SQL server 2012 |
| C6      | Django | Django-rest-framework, Django, Django Forms, Django models, Django-views, Django-templates |
| C7      | R      | R, Dplyr, Data Table, ggplot2,Tidyverse |
| C8      | Angular| Angular, Angular2-routing, Dart, Angular-dart, Angular JS, Angular6, Angular-material |
| C9      | Apple  | IOS, Swift, XCode, Swift4, UltableView |
| C10     | Excel  | Excel,Vba, Excel-formula, Excel-vba, Excel-2010 |
| C11     | OS     | Linux, Bash, shell, Unix, Macos, Sed |
| C12     | Asp.net| asp.net core, asp.net-mvc, asp.net, asp.net-mvc-4, razor |
| C13     | Entity Framework | Entity Framework, entity-framework-core,entity-framework-6 |
| C14     | C      | C,C++,C++14,Pointers, Struct, Objective-C, C++11,Scanf |
| C15     | JavaScript | JavaScript , JQuery, NodeJS, arrays, typescript, ReactJS, backbone-views, marionette, backbone.js |
| C16     | Git    | Git, Github, Eclipse, version-control, Merge, git-merge, Matrix |
| Tags  | Answerers                                      | No. of Experts | Expert Answerer |
|-------|-----------------------------------------------|----------------|-----------------|
| Bash  | A6, A8, A13, A14, A16, A67, A81              | 7              | A81             |
| Java  | A13, A28, A29, A33, A42, A43, A47, A49, A52, A60, A70, A86, A87 | 14             | A47             |
| Python-3.X | A16, A31, A39, A40, A44, A46, A48, A54, A70, A82, A85 | 11             | A46             |
| Numpy | A1, A13, A16, A39, A40, A44, A56              | 7              | A1              |
| CSS   | A4, A11, A12, A19, A24, A33, A55, A65, A72, A76 | 10             | A55             |
| Jquery| A4, A5, A7, A11, A12, A15, A18, A19, A22, A25, A41, A42, A65, A69, A72 | 15             | A42             |
| Pandas| A15, A39, A43, A47, A56, A64, A69, A84        | 8              | A39             |
| C#    | A5, A17, A23, A47, A51, A52, A53, A66, A69, A77, A78, A82 | 13             | A47             |
| Array | A1, A9, A11, A19, A22, A24, A35, A41, A42, A59, A62, A63, A66, A73, A83, A87 | 16             | A42             |
| Java Script | A4, A5, A7, A10, A11, A12, A15, A18, A19, A22, A23, A24, A25, A33, A41, A42, A50, A52, A55, A65, A69, A72, A76, A77, A79 | 25             | A42             |
| Html  | A2, A4, A5, A7, A10, A11, A12, A19, A22, A24, A25, A33, A41, A42, A49, A55, A65, A69, A72, A76 | 21             | A55             |
| Linux | A8, A13, A14, A60, A67, A73, A81              | 7              | A60             |
| C     | A13, A14, A27, A35, A59, A60, A73              | 7              | A37             |
| Angular js | A2, A10, A22, A23, A28, A33, A50, A57, A77 | 9              | A50             |
| Python | A1, A4, A13, A16, A20, A30, A31, A37, A39, A40, A44, A46, A48, A52, A54, A56, A57, A59, A64, A67, A70, A74, A75, A81, A82, A85 | 26             | A46             |
| Regex | A3, A5, A21, A45, A52                          | 5              | A52             |
| C++   | A5, A9, A13, A14, A26, A27, A37, A51, A58, A60, A71, A73 | 12             | A51             |
Table 3 shows the identified experts for the given question by using our proposed system. We have evinced the results of 10 questions in table 3. As a first step, from the given question, the keyword is extracted by using question pre-processing. For the identified keyword, the cluster is identified by comparing the keyword and the tag in each cluster. If the tag is matched, then the expert is recognized from the cluster.

**Table 3. Identified experts for the given question**

| Question                          | Keyword        | Cluster name | Expert name     |
|----------------------------------|----------------|--------------|-----------------|
| What is python-2.7               | Python-2.7     | Python       | Akrun           |
| Features of django               | Django         | Django       | Torek           |
| What is meant by list in python  | List ,python   | Python       | Akrun           |
| What is generics                 | Generics       | .net         | Marktijnipieter |
| Brief about sql server           | Sql server     | Sql          | Unwind          |
| Advantages of using word press   | Wordpress      | Web          | Scaisedge       |
| Define tidy verse                | Tidy verse     | R            | Wen             |
| Difference between qt and qt5    | Qt,qt5         | Qt           | Argo            |
| Explain database design          | Database design| Database     | Paxdiablo       |
| Define algorithm                 | Algorithm      | Others       | John zwinck     |
In every cluster, there are several tags, and the corresponding expert is stored in the form of key-value pairs. Fig. (3) is the pictorial representation of the number of identified experts and tags in each cluster.

![Graph showing experts and tags](image)

**Figure 3.** Experts and Tags in each cluster

### 6. Conclusion

This paper recommends the expert for the given question by using the cluster approach. In CQA websites, the user posts the question and gets precise and relevant answers. This paper proposed a cluster-based approach for finding experts from CQA website Stack Overflow. Our proposed work involves the process such as question preprocessing, tag identification, cluster analysis, and expert identification. Firstly, domain-wise clusters are created by doing the tag analysis process. Secondly, the keywords are extracted from the given question by doing data preprocessing techniques. The experts are identified by doing string matching between extracted keyword and tags exist in the cluster. A similarity score is computed between the keyword and the experts in a particular domain. Based on the score, the expert is identified from the domain. The person who has the highest score is to be the expert in that domain.

### References

[1] Abdelghani Bouziane, Djelloul Bouchiha, Noureddine Doumi and Mimoun Malki 2015 Question Answering Systems: Survey and Trends *Procedia Computer Science*. 73 pp 366 – 375. https://doi.org/10.1016/j.procs.2015.12.005

[2] Annamoradnejad I, Fazli M and Habibi J 2020 Predicting Subjective Features from Questions on QA Websites using BERT *Proc. of 16th Int. Conf. on Web Research (Iran)* pp. 240-244. https://doi.org/10.1109/ICWR49608.2020.9122318.

[3] Antoaneta Baltadzhieva and Grzegorz Chrupala 2015 Question Quality in Community Question Answering Forums: a Survey *ACM SIGKDD Explorations Newsletter*. 17(1) pp 8-13. https://doi.org/10.1145/2830544.2830547

[4] Baichuan Li and Irwin King, 2010 Routing questions to appropriate answerers in community question and answer services *ACM*, pp 1585-1588. https://doi.org/10.1145/1871437.1871678
Dallia Elalfy, Walaa Gad and Rasha Ismail 2018 A Hybrid Model To Predict Best Answers In Question Answering Communities Egyptian Informatics Journal. 19(1) pp 21-31. https://doi.org/10.1016/j.eij.2017.06.002

Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei and Evangelos Milios, 2012, “Finding expert’s users in community question answering”, Proc. of the 21st Int. Conf. on World Wide Web pp. 791-798 https://doi.org/10.1145/2187980.2188202

Hapnes Toba, Zhao-Yan Ming, Mirna Adriani and Tat-Seng Chua 2014 Discovering high-quality answers in community question answering archives using a hierarchy of classifiers Information Sciences, Elsevier 261 pp 101-15 https://doi.org/10.1016/j.ins.2013.10.030

Imrul Kayes, Nicolas Kourtellis and Adriana Iamnitchi, 2017 The Good, The Bad and The Deviant In Community Question Answering Online Social Networks and Media, Elsevier 2 pp 45-59 https://doi.org/10.1016/j.osnem.2017.08.002

Maheshwari K B , Nandhini Priya J, Shanthaa Lakshmi M and Menaha R 2018 Question Answering System For Voice Based Search Using NLP Techniques And Web Snippets International Journal of Scientific & Engineering Research, 9(3) ISSN 2229-5518.

Maria Soledad Pera and Yiu-Kai Ng 2011 A Community Question-Answering Refinement System", Proc. of the 22nd ACM Conf. on Hypertext and Hypermedia, Eindhoven, (Netherlands), https://doi.org/10.1145/1995966.1995999.

Menaha R, Udhaya Surya A, Nandhini K and Ishwarya M 2017 Question answering system using web snippets Int. Conf. on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (Palladam) pp 387-390 https://doi.org/10.1109/I-SMAC.2017.8058377

Mishra A and Jain SK. 2016 A survey on question answering systems with classification J. King Saud Univ Comput Inf Sci 28(3) pp 345–361. https://doi.org/10.1016/j.jksuci.2014.10.007

Roy, P K , Ahmad, Z and Singh, J P 2018 Finding and Ranking High-Quality Answers in Community Question Answering Sites Glob J Flex Syst Manag. 19 pp 53–68. https://doi.org/10.1007/s40171-017-0172-6

Wenhao Zheng and Ming Li 2017 The Best Answer Prediction By Exploiting Heterogeneous Data On Software Development Q&A Forum Neurocomputing , Elsevier. 269 pp. 212-219. https://doi.org/10.1016/j.neucom.2016.12.097.

Yuan Yao, Hanghang Tong, Tao Xie, Leman Akoglu, Feng Xu, and Jian Lu, 2015 Detecting High-Quality Posts In Community Question Answering Sites, Information Science, Elsevier 302 pp 70-82. https://doi.org/10.1016/j.ins.2014.12.038

Zhenlei Yan and Jie Zhou 2015 Optimal Answerer Ranking For New Questions In Community Question Answering Information Processing and Management, Elsevier 51(1) pp 163-178. https://doi.org/10.1016/j.ipm.2014.07.009