# Generating Appropriate Question-Answer Pairs for Chatbots using Data Harvested from Community-based QA Sites

Wenjing Yang and Jie Wang
*Department of Computer Science, University of Massachusetts, Lowell, MA 01854, U.S.A.*

Keywords: Question-Answering, Content Extraction, LDA Clustering, Keyword Extraction.

Abstract: Community-based question-answering web sites (CQAW) contain rich collections of question-answer pages, where a single question often has multiple answers written by different authors with different aspects. We study how to harvest new question-answer pairs from CQAWs so that each question-answer pair addresses just one aspect that are suitable for chatbots over a specific domain. In particular, we first extract all answers to a question from a CQAW site using DOM-tree similarities and features of answer areas, and then cluster the answers using LDA. Next, we form a sub-question for each cluster using a small number of top keywords in the given cluster with the keywords in the original question. We select the best answer to the sub-question based on user ratings and similarities of answers to the sub-question. Experimental results show that our approach is effective.

## 1 INTRODUCTION

Community-based question-answering web sites (CQAWs) are rich depositories of information used by many people. For example, Zhihu, the most popular Chinese CQAW has attracted over 65 million users since 2011. Quora, the most popular English CQAW, has attracted over 120 millions users since 2009. CQAW sites allow users to ask any question over any subject. To control the quality, only domain experts are supposed to provide answers to user questions. This requirement, unfortunately, is not guaranteed as in any public forum. Nevertheless, CQAWs are still rich sources of data in the form of question followed by multiple answers. In particular, answers provided by different authors to a given question often reflect different aspects of the question. As most questions allow for different perspectives, a human answering the question would choose one of the possible interpretations and provide an appropriate answer.

This motivates us to harvest question-answering data from CQAW sites and generate new question-answer pairs (QAPs) so that they are suitable for chatbots. A chatbot is a conversation agent, which mimics humans to communicate with users using natural languages.

To achieve this, we first extract questions and their answers from CQAWs. But the data extracted this way often cannot be used directly by chatbots, for a question may have a large number of answers addressing different aspects. Chatbots should not return multiple answers as search engines typically do.

We have two tasks: (1) Extract questions and their answers from CQAWs. (2) Construct new QAPs such that the answer to the question in each QAP is correct, is reasonably short, and addresses only one aspect.

To achieve the first goal, we investigate the layouts of the most popular CQAW sites, based on which we use HTML tag similarities of the DOM subtrees and two HTML features of answer areas to identify and extract answers. The first feature is the number of `<DIV>` tags contained in an answer, and the second is the difference of the numbers of `<DIV>` tags between any two answers. The first feature is determined by a large lower bound and the second by a small upper bound. Based on these we construct a QA Extractor (QAE). QAE first identifies the questions from a given question page, and the answer area from a corresponding answer page linked from the question. It then extracts questions and all the answers to the corresponding question using text densities (Wang and Wang, 2015) and regular expressions. Experiments show that our QAE system achieves high accuracies. In particular, among the six popular CQAWs we tested, four have F1 scores from 95% to 100%, one (Quora) has 85% and one (Sina) has 90%.

To achieve the second goal, for each question $q$

and the set $S_q$ of all answers to $q$, we use LDA (Blei et al., 2003) to obtain a clustering of $S_q$. Moreover, we compute an appropriate number $K_q$ of clusters for each question $q$, so that each cluster has one focus with a reasonable size. We then form a new sub-question for each cluster using the keywords in the original question $q$ plus a few top keywords in the cluster. We select an answer in the cluster such that it has the highest user rating and the highest cosine similarity to the sub-question. The sub-question and the chosen answer form a new QAP. Human evaluations in our experiments show that, after removing spam contents, 79% of the new QAPs make sense. The reason some QAPs do not make sense is that authors of the answers often include something in their answers that are not really relevant to the question. When these answers are clustered to form QAPs, the answers would seem irrelevant to the original question. Finally, we develop a question-answer-pair system (QAPS) and an app to display QAPs. We tested the QAPs over the domain of diabetes we harvested from CAQWs on a chatbot devised by Zhang and Wang (Zhang and Wang, 2017); the results are satisfactory.

The rest of the paper is organized as follows: We describe related work in Section 2 and present our system architecture in Section 3. We present our algorithms to extract questions and the corresponding answers from CQAW sites in Section 4, and construct QAPs in Section 5. We show our experiment results in Section 6 and conclude the paper in Section 7.

## 2 RELATED WORK

Content extraction from web pages has been studied extensively, resulting in a variety of solutions. Text densities (Wang and Wang, 2015) and edit distance (Reis et al., 2004), for example, are two successful content-extraction methods for news websites.

We note that in an answer page, answers have similar layouts. But other non-answer areas such as navigation bars and lists of recommendations may also have repetitive patterns. Nevertheless, to extract exactly the answers from CQAW sites, we first need to identify repetitive pattern regions. We then use unique features of answer areas to identify answers. Repetitive pattern regions can be identified by HTML tag sequences such as IEPAD (Chang and Lui, 2001) and Dela (Arasu and Garcia-Molina, 2003). MDR (Liu et al., 2003) is another method that locates repetitive regions based on the DOM tree of the page. These methods adopt edit distance to determine which two regions are similar. While these methods may be used for our purposes, we note that the structures of CQAW pages often topic recommendations, friend recommendations, and other types of noisy data. Using these methods may lead to poor extractions of contents.

Machine learning methods have also been investigated, some of which uses different templates to learn the most important part of a web page (Yang et al., 2009) and some of which uses SVM (Support Vector Machine) to perform shallow information extraction (Yu et al., 2002). Machine learning methods, however, are time-consuming with the problem of cold start.

Most chatbots have a dialog module that manages the conversation process and determines what to reply to the question asked by the user. Early research paid little attention to answers that cover different aspects of the question. We note that one question with multiple answers that address different aspects may be harvested to generate multiple QAPs suitable for chatbots. Therefore, chatbots would benefit from high-quality QAPs as raw materials, where each QAP only addresses one aspect. Clustering answers to a question would be a good start for constructing such QAPs.

Ranking SVM is a useful method to construct QAPs based on answer qualities (Huang et al., 2007). On the other hand, LDA (Latent Dirichlet Allocation), after modifications, was shown to offer better clustering results on shorter text messages (Jin et al., 2011) for constructing QAPs. HDP (Hierarchical Dirichlet Process)-LDA (Teh et al., 2005) is a effective method to compute the number of clusters, which works well on large numbers of clusters. Based on our experiments, the number of clusters for answers to a question is quite small. Thus, we choose LDA instead of SVM to compute clusterings, and we devise a new method to compute the number of clusters instead of using HDP.

Sequential rules and non-textual features based on CQAW structures may also be used to determine QAPs (Wang et al., 2009). Machine learning method have also been used in ranking a answer (Surdeanu et al., 2008). However, these method may not consider the review of other users, which is a important feature of CQAWS.

## 3 ARCHITECTURE

The architecture of our system, shown in Figure 1, consists of four components: (1) QAE; (2) AC (Answer Clustering); (3) QAPG (QAP Generation); (4) QAD (QA Display).

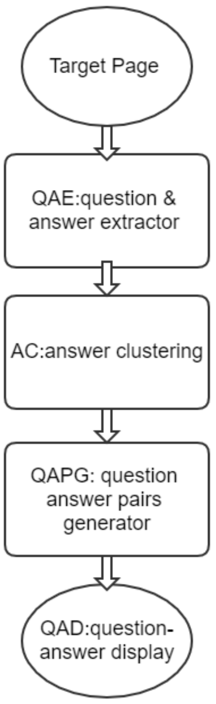

Target Page
QAE:question & answer extractor
AC:answer clustering
QAPG: question answer pairs generator
QAD:question-answer display

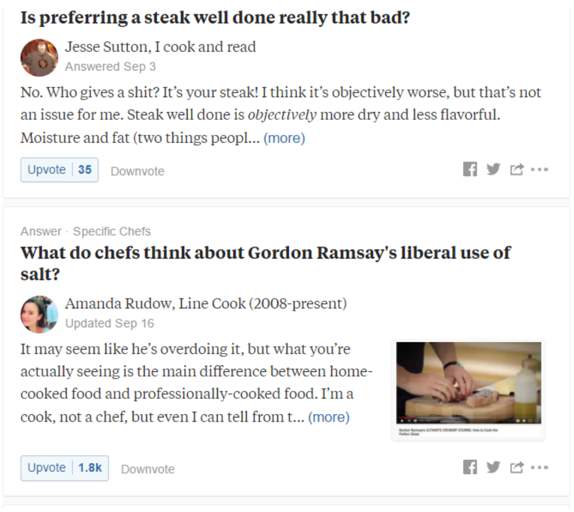
Figure 1: System architecture.

Given below are brief descriptions of these components:

- QAE extracts questions, all the answers to the corresponding questions, and user rating of an answer. These are our raw data for QAPs.
- AC divides, for each question $q$, all the answers to $q$ into $K_q$ clusters, where a cluster represents a particular focus. The value of $K_q$ depends on the answers to $q$, which may be different for different questions.
- QAPG generates $K_q$ QAPs, where each cluster contributes exactly one QAP.
- QAD is an app that displays, for each question $q$, all the QAPs based on $q$.

## 4 QAE ALGORITHM

QACW sites typically consist of question pages and answer pages. A question page displays an enumeration of questions. Each question is linked to an answer page, which displays all the answers to the question, with some other information for each answer, such as the author's name who wrote the answer, the time when the answer was posted, and readers' ratings, among other things. Figure 2 (a) and (b) depict, respectively, a sample question page and a sample answer page.



**Is preferring a steak well done really that bad?**

Jesse Sutton, I cook and read
Answered Sep 3

No. Who gives a shit? It's your steak! I think it's objectively worse, but that's not an issue for me. Steak well done is *objectively* more dry and less flavorful. Moisture and fat (two things peopl... (more)

Upvote 35 Downvote

Answer · Specific Chefs

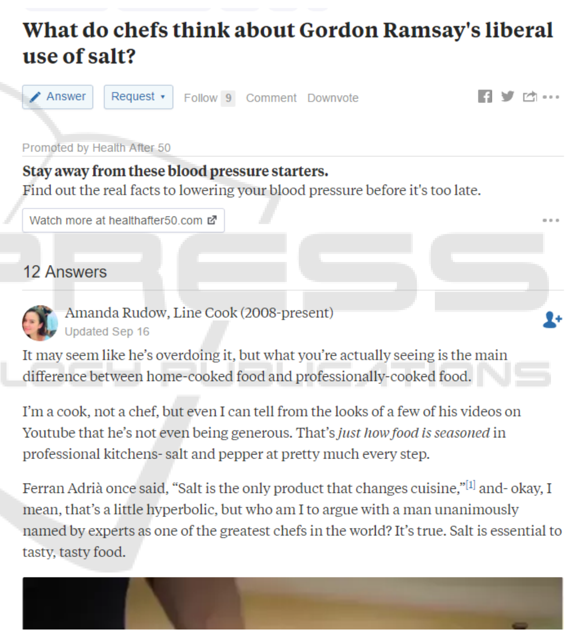
**What do chefs think about Gordon Ramsay's liberal use of salt?**

Amanda Rudow, Line Cook (2008-present)
Updated Sep 16

It may seem like he's overdoing it, but what you're actually seeing is the main difference between home-cooked food and professionally-cooked food. I'm a cook, not a chef, but even I can tell from t... (more)

Upvote 1.8k Downvote

(a)



**What do chefs think about Gordon Ramsay's liberal use of salt?**

Answer Request Follow 9 Comment Downvote

Promoted by Health After 50

**Stay away from these blood pressure starters.**

Find out the real facts to lowering your blood pressure before it's too late.

Watch more at healthafter50.com

12 Answers

Amanda Rudow, Line Cook (2008-present)
Updated Sep 16

It may seem like he's overdoing it, but what you're actually seeing is the main difference between home-cooked food and professionally-cooked food.

I'm a cook, not a chef, but even I can tell from the looks of a few of his videos on Youtube that he's not even being generous. That's *just how food is seasoned* in professional kitchens- salt and pepper at pretty much every step.

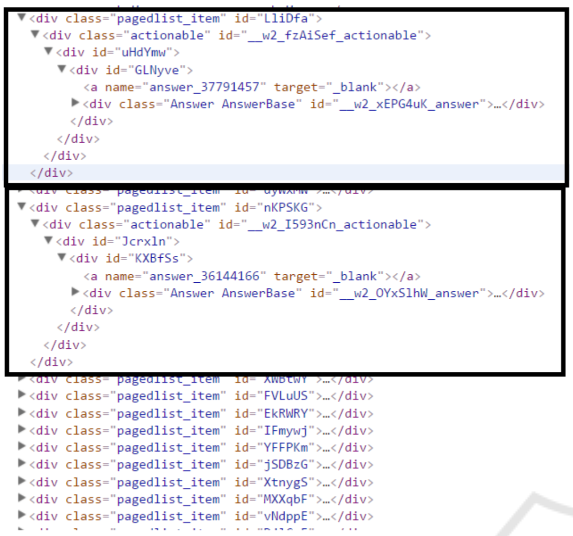
Ferran Adrià once said, "Salt is the only product that changes cuisine,"[1] and- okay, I mean, that's a little hyperbolic, but who am I to argue with a man unanimously named by experts as one of the greatest chefs in the world? It's true. Salt is essential to tasty, tasty food.

(b)

Figure 2: (a) Sample question page. (b) Sample answer page.

Extracting questions from a question page is straightforward. To extract answers, we must first identify where the answers are located on an answer page.

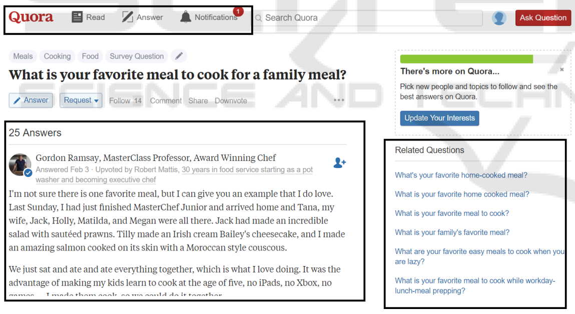
Most CQAW sites use the same template to generate HTML pages, which have similar layouts. From Figure 2 (a) we can see that the HTML-tag layouts of the two rectangles for displaying answers are exactly the same. For convenience, we will refer to an area for displaying answers as the answer area. Figure 3 depicts the DOM-tree structure of an answer page, from which we can see that the DOM subtrees in answer areas have similar layouts, and they differ from the other parts of the page.



```
▼<div class="pagedlist_item" id="LliDfa">
  ▼<div class="actionable" id="__w2_fzAiSef_actionable">
    ▼<div id="uHdYmw">
      ▼<div id="GLNyve">
          <a name="answer_37791457" target="_blank"></a>
        ▶<div class="Answer AnswerBase" id="__w2_xEPG4uK_answer">…</div>
        </div>
      </div>
    </div>
  </div>
▼<div class="pagedlist_item" id="nKPSKG">
  ▼<div class="actionable" id="__w2_I593nCn_actionable">
    ▼<div id="Jcrxln">
      ▼<div id="KXBfSs">
          <a name="answer_36144166" target="_blank"></a>
        ▶<div class="Answer AnswerBase" id="__w2_OYxSlhW_answer">…</div>
        </div>
      </div>
    </div>
  </div>
▶<div class="pagedlist_item" id="FVLuUS">…</div>
▶<div class="pagedlist_item" id="EkRWRY">…</div>
▶<div class="pagedlist_item" id="IFmywj">…</div>
▶<div class="pagedlist_item" id="YFFPKm">…</div>
▶<div class="pagedlist_item" id="jSDBzG">…</div>
▶<div class="pagedlist_item" id="XtnygS">…</div>
▶<div class="pagedlist_item" id="MXXqbF">…</div>
▶<div class="pagedlist_item" id="vNdppE">…</div>
```

Figure 3: DOM tree over the answer area.

A typical answer page contains a list of answers and other information such as navigation bars and recommendations (see Figure 4).



What is your favorite meal to cook for a family meal?

There's more on Quora...

Pick new people and topics to follow and see the best answers on Quora.

25 Answers

Gordon Ramsay, MasterClass Professor, Award Winning Chef
Answered Feb 3 · Upvoted by Robert Mattis, 30 years in food service starting as a pot washer and becoming executive chef

I'm not sure there is one favorite meal, but I can give you an example that I do love. Last Sunday, I had just finished MasterChef Junior and arrived home and Tana, my wife, Jack, Holly, Matilda, and Megan were all there. Jack had made an incredible salad with sautéed prawns. Tilly made an Irish cream Bailey's cheesecake, and I made an amazing salmon cooked on its skin with a Moroccan style couscous.

We just sat and ate and ate everything together, which is what I love doing. It was the advantage of making my kids learn to cook at the age of five, no iPads, no Xbox, no

Related Questions

What's your favorite home-cooked meal?
What is your favorite home cooked meal?
What is your favorite meal to cook?
What is your family's favorite meal?
What are your favorite easy meals to cook when you are lazy?
What is your favorite meal to cook while workday-lunch-meal prepping?

Figure 4: A typical answer page.

Most CQAW sites use nested `<DIV></DIV>` tags to represent an answer area, and in the most inner `<DIV></DIV>` layer contains content tags such as `<P></P>` and `<IMG></IMG>`. Moreover, answers are listed in parallel at the same level in the DOM tree, and each answer is enclosed in a `<DIV></DIV>` tag pair. Assume that the answer area starts at level $N$. In the DOM tree of an answer page, the answer area is in a subtree with the largest number of `<DIV>` tags, rooted at a `<DIV>` tag at level $N-1$, and the answers are listed at level $N$, which have similar structures pairwise (see Figure 5).

We represent each DOM subtree as a string of HTML tags by concatenating all the tag nodes level by level from top to bottom, and at each level from left

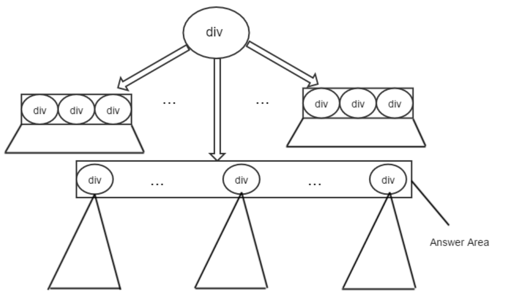


Figure 5: Answer area.

to right. We say that two subtrees have similar structures if the cosine similarity of the two corresponding tag strings exceeds a threshold. In our experiments, an empirical value of this threshold is set to 0.9.

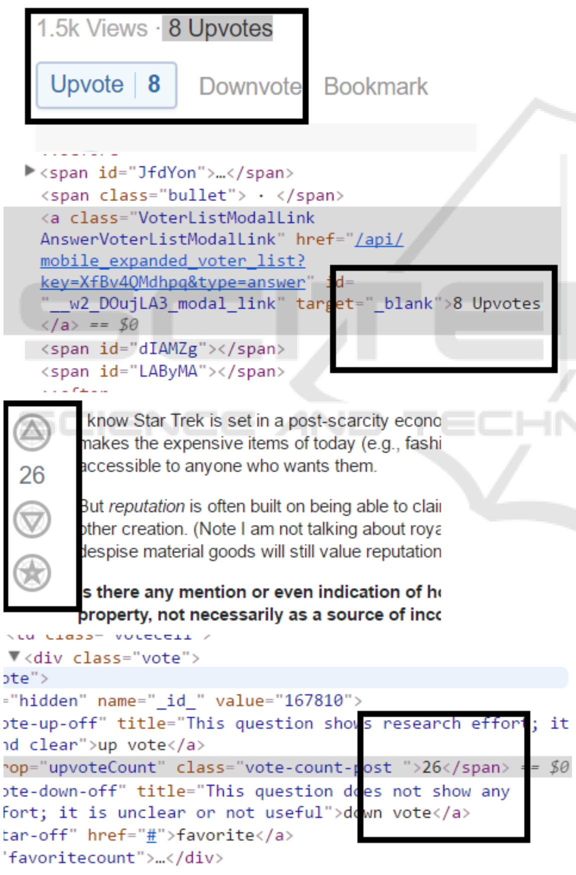
Moreover, let NDA denote the number of `<DIV>` tags contained in an answer-subtree, then NDA for any answer has a nontrivial lower bound $M$. Let DDAP denote the absolute value of the difference between two NDAs, then DDAP for any pair of answers has a small upper bound $D$.

We note that in a special case with only one `<DIV>` tag at level $N$, either there is only one answer or there is no answer at all at this point. As we use repeated pattern to detect the answer area, our algorithm fails to identify the answer (if there is any) in this case. To overcome this drawback, we will just need to crawl the page again when more answers are posted, which will happen soon if a question attracts sufficient attention. We note that popular questions often have two or more answers. In what follows, we assume that there are at least two answers in the answer area.

Let $T$ be the DOM tree of an answer page. We identify answers recursively over $T$ as follows: Start from $N = 1$, find the subtree $T'$ at this level with the largest number of `<DIV>` tags. In subtree $T'$, look at all subtrees at level $N+1$, and check if there are at least two subtrees having similar structures, and each of these subtrees contains at least $M$ `<DIV>` tags. and the absolute difference of the numbers of `<DIV>` tags of the two subtrees is at most $D$. We set an empirical value of $M$ to 10 and an empirical value of $d$ to 5 in our experiments. If so, then compare all subtrees at this level pairwise and collect those that have similar structures such that the corresponding NDA is at least $M$ and the corresponding DDAP is at most $D$. This collection is the extracted answer area. If not, then set $N$ to $N+1$ and repeat the above procedure.

After identifying the answer area, we extract the answers (text) and the rating (number, if there is any) of each answer. To extract answers, we will use the text-density method devised in (Wang and Wang, 2015) by treating each subtree in the answer area as a web page by itself. We may also write regular expressions to identify the answer area and extract answers. This approach depends on the structure of a particular CQAW site, and it is difficult to write a uniform regular expression suitable for all CQAW sites.

The rating of an answer is a feature of CQAW sites, which is an indicator of the quality of the answer. CAQWs sites typically list the answer with the highest rating on the top of an answer page. From Figure 6, we can see different CQAW sites' rating styles. After analyzing a lot of answer areas from different web sites, for each CQAW we write a regular expression to extract the ratings. For example, for the first answer page shown on the top of Figure 6, the regular expression is `^>(.*)Upvotes$`, and for the second answer page the regular expression is `^vote-count-post">(.*)</span$`.



```
▶<span id="JfdYon">…</span>
 <span class="bullet"> · </span>
 <a class="VoterListModalLink
 AnswerVoterListModalLink" href="/api/
 mobile_expanded_voter_list?
 key=XfBv4QMdhpq&type=answer" id=
 "__w2_DOujLA3_modal_link" target="_blank">8 Upvotes
 </a> == $0
 <span id="dIAMZg"></span>
 <span id="LAByMA"></span>
```

know Star Trek is set in a post-scarcity econo
makes the expensive items of today (e.g., fashi
accessible to anyone who wants them.

But *reputation* is often built on being able to clai
other creation. (Note I am not talking about roya
despise material goods will still value reputation

**s there any mention or even indication of h**
**property, not necessarily as a source of inc**

```
▼<div class="vote">
ote">
="hidden" name="_id_" value="167810">
ote-up-off" title="This question shows research effort; it
nd clear">up vote</a>
rop="upvoteCount" class="vote-count-post ">26</span> == $0
ote-down-off" title="This question does not show any
fort; it is unclear or not useful">down vote</a>
tar-off" href="#">favorite</a>
'favoritecount">…</div>
```

Figure 6: Sample ratings and DOM subtrees of different answers.

## 5 AC AND QAPG ALGORITHMS

After extracting the set of answers to a question, we use Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to obtain a clustering of the answers with a dynamically determined number of clusters.

### 5.1 LDA

LDA (Blei et al., 2003) is a probabilistic generative model, treating each document $d$ having a mixture of $K$ latent topics, where each topic $z$ is modeled by a probability distribution over words in the vocabulary from a corpus of documents including $d$. LDA proceeds as follows:

For each word $w_n$ in document $d$:

1. Draw a topic distribution $z_n \sim Multinomial(\boldsymbol{\theta})$ with a hyper parameter $\boldsymbol{\alpha}$.
2. Draw a word distribution $w_n \sim Multinomial(\boldsymbol{\phi}_{Z_n})$ with a hyper parameter $\boldsymbol{\beta}$.

Here $Multinomial(\boldsymbol{X})$ with parameter $\boldsymbol{p}$ represents a multinomial distribution of hidden random variable $\boldsymbol{X}$. It is customary to use Gibbs sampling estimation to compute a document-topic probability matrix and a topic-word probability matrix. Each topic corresponds to a cluster. For topic $k$, the set of documents that have the highest probabilities under $k$ forms a cluster of topic $k$.

### 5.2 Compute the Number of Clusters

We treat each answer in the set of extracted answers to a question as a document, and apply LDA to cluster the answers. To do so, we need to determine an appropriate value of $K$ using the following algorithm. Initially set $K = 2$.

1. Run LDA on the set of answers to generate $K$ clusters.
2. If there is an empty cluster, then $K-1$ is good and stop.
3. Select $W$ top keywords from each cluster; namely, the first $W$ keywords with the highest probabilities under the topic corresponding to the cluster. For each pair of clusters $C_i$ and $C_j$, if the cosine similarity of the top $W$ keywords of $C_i$ and the top $W$ keywords of $C_j$ is less than a threshold $t_1$, then $K$ is good and stop. In other words, we have obtained a clustering where each cluster has a different aspect.
4. If there is a cluster such that the cosine similarity of the keywords in each answer in the cluster and the keywords of the cluster is smaller than a threshold $t_2$, then $K-1$ is good and stop. In other words, when the set of answers is divided into $K-1$ clusters, each cluster contains an answer that is similar to the cluster.

5. Otherwise, increase $K$ by 1 and repeat the algorithm while $K \leq L$ for some value of $L$.

Empirically we choose $L = 50$ and $W = 10$.

## 5.3 Generating QAPs

After clustering the answers, each cluster should center on just one aspect, and so the original question would be too broad for this cluster. We would like to generate a sub-question for each cluster and a short answer to the sub-question to form a meaningful QAP.

We remark that for a given a database of QAPs, how to find the most appropriate answer to a user question from the database is nontrivial; the obvious approach of using the best matching of the user question and an existing question in QAPs to find an answer is problematic. A novel approach to this problem will be presented in a separate paper.

### 5.3.1 Generate a Sub-question

In a QA system, a question is often represented by the keywords contained in it (after removing stop words). To generate a sub-question for a cluster, we use the TextRank (Mihalcea and Tarau, 2004) algorithm to rank keywords and add the top $R$ keywords to the set of keywords in the original question to form a sub-question. An empirical value of $R$ is 6.

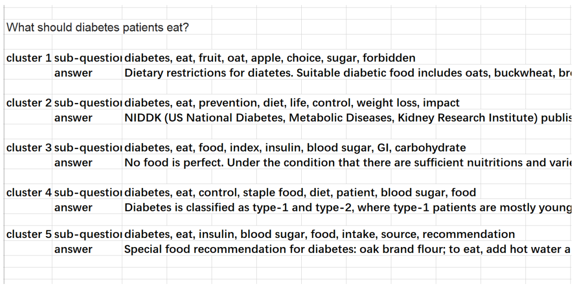
### 5.3.2 Choose an Answer

After clustering, we may still have multiple answers in a cluster. Since answers often have user ratings, we use user ratings and cosine similarity of an answer and the sub-question to select the best answer to the sub-question as follows:

1. Select answers with the highest user rating and place them in a set $U$.
2. Select answers with the highest cosine similarity with the sub-question, and place them in a set $V$.
3. Choose an answer as follows: If $U \cup V \neq \emptyset$, select an answer from the intersection. Otherwise, if the rating for $U$ is less than or equal to a threshold $T$, then we select an answer from $V$ to avoid fake ratings. If the rating for selecting $U$ is greater than $L$, then we select an answer from $U$. An empirical value of $T$ is set to 5 in our experiments.

## 5.4 Displaying

We develop a QAD app to display QAPs. Figure 7 shows an example of QAPs generated by our system.



What should diabetes patients eat?

| | | |
|---|---|---|
| cluster 1 | sub-questio | diabetes, eat, fruit, oat, apple, choice, sugar, forbidden |
| | answer | Dietary restrictions for diatetes. Suitable diabetic food includes oats, buckwheat, br |
| cluster 2 | sub-questio | diabetes, eat, prevention, diet, life, control, weight loss, impact |
| | answer | NIDDK (US National Diabetes, Metabolic Diseases, Kidney Research Institute) publis |
| cluster 3 | sub-questio | diabetes, eat, food, index, insulin, blood sugar, GI, carbohydrate |
| | answer | No food is perfect. Under the condition that there are sufficient nuitritions and vari |
| cluster 4 | sub-questio | diabetes, eat, control, staple food, diet, patient, blood sugar, food |
| | answer | Diabetes is classified as type-1 and type-2, where type-1 patients are mostly young |
| cluster 5 | sub-questio | diabetes, eat, insulin, blood sugar, food, intake, source, recommendation |
| | answer | Special food recommendation for diabetes: oak brand flour; to eat, add hot water a |

Figure 7: Sample answer clustering for the question: "What should diabetes patients eat?".

# 6 EXPERIMENTS

Experiments were carried out on a desktop with a 3.5 GHZ Intel Core i7, 16GB DDR3, and Windows 10 operation system. We wrote our programs in Java, executed with a single thread.

We crawled 500 pages from 10 CQAW sites covering different topics (see Table 1).

Table 1: 10 CQAW sites.

| **Chinese CQAW** | **URL** |
|---|---|
| 360question | http://wenda.so.com/ |
| zhihu | https://www.zhihu.com |
| baidu zhidao | http://zhidao.baidu.com |
| tianya | http://wenda.tianya.cn/ |
| sina ask | http://iask.sina.com.cn/ |
| **English CQAW** | |
| quora | https://www.quora.com/ |
| stackoverflow | https://stackoverflow.com/ |
| stack exange | https://stackexchange.com/ |
| yahoo answers | https://answers.yahoo.com/ |
| blurt it | http://www.blurtit.com/ |

## 6.1 Extraction Accuracy and Analysis

We chose 50 answer pages at random from each CQAW site. Let TNA denote the total number of answers included in these pages (the true number) and ST the similarity threshold to determine if two subtrees at level $N$ are similar. Recall that $M$ denotes the lower bound of the number of `<DIV>` tags in an answer and $D$ the upper bound of the absolute difference between the numbers of `<DIV>` tags of any two answers. We measure the accuracy of extractions as follows:

$$\begin{aligned} Precision &= \frac{ACE}{AER}, \\ Recall &= \frac{ACE}{ACR}, \\ F_1 &= \frac{2 \times Precision \times Recall}{Precision + Recall}, \end{aligned}$$

where $ACE$ denotes the number of correctly extracted answers, $AER$ the total number of all extracted results, and $ACR$ the total number of all answers. Experiment results indicate that when choosing ST = 0.7, $D = 5$, and $M = 10$, we achieved the highest $F_1$-scores over both Chinese and English pages (see Table 3).

Table 2: Extraction accuracy on datasets.

| | **precision** | **recall** | **$F_1$ scores** |
|---|---|---|---|
| Chinese CQAWs | 100% | 93% | 96% |
| English CQAWs | 96.50% | 91% | 93% |

When answer A includes more images or other types of content over answer B, the answer area of A will contain more `<DIV>` tags, resulting in a higher value of $D$. We note that setting a higher value of $M$, we may filter out more noise, but simpler answer may have a smaller value of $M$. To make the system more accurate, users may modify the values of $D$ and $M$ to find better values.

Table 3: Extraction accuracy.

| **CQAW** | **precision** | **recall** | **$F_1$ score** |
|---|---|---|---|
| 360 | 100% | 90.50% | 95% |
| zhihu | 100% | 92% | 96% |
| baidu | 100% | 100% | 100% |
| sina | 100% | 81.50% | 90% |
| quora | 91% | 80% | 85% |
| stackoverflow | 100% | 100% | 100% |

## 6.2 Clustering

We choose at random a number of questions from the domain of diabetes. Let ANA denote the average number of answers for each question. Denote by CS the threshold upper bound of cosine similarity between different clusters over the top 10 keywords in each cluster, and ASC the threshold upper bound of the cosine similarity between an answer and the cluster the answer belongs to over the top 10 keywords in an answer and a cluster, respectively. Different values of CS and ACS would result in a different number $K$ of clusters.

We compute the number of clusters for each answer page by requiring that no cluster is empty. The results are shown in Table 4.

## 6.3 Quality of QAPs

We randomly choose 50 questions covering different topics in the domain of diabetes and generate new QAPs. We ask a number of volunteers with expertise in the underlying domain to label the quality of a QAP "good" if the QAP makes sense, "spam" if the QAP clearly contains spam contents, and "bad" if the QAP does not make sense. The results are shown in Table 5. Our diabetes QAPs data have been used in a chatbot presented in (Zhang and Wang, 2017).

Table 4: Topic number in English and Chinese dataset with different parameters.

| **CQAW** | **ANA** | **CS** | **ACS** | **AK** |
|---|---|---|---|---|
| Chinese | 105 | 0.20 | 0.13 | 3.2 |
| | | 0.15 | 0.13 | 5.5 |
| | | 0.13 | 0.13 | 7.2 |
| | | 0.10 | 0.13 | 8.5 |
| | | 0.08 | 0.13 | 8.4 |
| | | 0.20 | 0.18 | 3.2 |
| | | 0.15 | 0.18 | 5.8 |
| | | 0.13 | 0.18 | 6.6 |
| | | 0.10 | 0.18 | 7.2 |
| | | 0.08 | 0.18 | 7.8 |
| | | 0.20 | 0.24 | 3.2 |
| | | 0.15 | 0.24 | 4.5 |
| | | 0.13 | 0.24 | 5.8 |
| | | 0.10 | 0.24 | 6.8 |
| | | 0.08 | 0.24 | 7.5 |
| English | 122 | 0.20 | 0.13 | 3.0 |
| | | 0.15 | 0.13 | 5.6 |
| | | 0.13 | 0.13 | 7.2 |
| | | 0.10 | 0.13 | 7.8 |
| | | 0.08 | 0.13 | 8.0 |
| | | 0.20 | 0.18 | 3.0 |
| | | 0.15 | 0.18 | 5.6 |
| | | 0.13 | 0.18 | 6.0 |
| | | 0.10 | 0.18 | 7.0 |
| | | 0.08 | 0.18 | 7.6 |
| | | 0.20 | 0.24 | 3.0 |
| | | 0.15 | 0.24 | 5.2 |
| | | 0.13 | 0.24 | 6.0 |
| | | 0.10 | 0.24 | 6.6 |
| | | 0.08 | 0.24 | 7.0 |

Table 5: Evaluation results by humans.

| **Number of new QAPs** | **good** | **spam** | **bad** |
|---|---|---|---|
| 410 | 304 | 25 | 81 |
| 385 (excluding spams) | 79% | | 21% |

Spam contents are common in CQAWs. Detecting spam contents, not addressed in this paper, will be investigated in a future project. We also note that the main reason for the QAPs to be labeled "bad" is that most of these answers, even though they are not spams, are less relevant to the original question. Thus, the quality of the original question-answering data is important. One way to improve the ratio of good QAPs is to dissociate the original question from these clusters.

## 7 CONCLUSION AND FUTURE WORK

We presented a novel method to extract the rich question-answering information from community-based QA web sites and generate a large number of QAPs. Experiments confirm that our method has a high accuracy of extracting information and a high quality of constructing QAPs suitable for chatbots.

We plan to investigate the following issues in future studies: (1) Improve the ratio of good QAPs by dissociating certain clusters with the original question. (2) Generate an aggregate answer from different answers in a cluster to a sub-question, instead of just choosing the best one. (3) Devise a spam detection mechanism to filter out spams from the data collected from CQAW sites.(4) Compare the number of clusters generated by HDP-LDA with the number of clusters generated by our algorithm.

## ACKNOWLEDGEMENTS

This work was supported in part by Eola Solutions Inc. The authors are grateful to members of the Text Automation Lab at UMass Lowell for discussions.

## REFERENCES

Arasu, A. and Garcia-Molina, H. (2003). Extracting structured data from web pages. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 337–348. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Chang, C.-H. and Lui, S.-C. (2001). Iepad: Information extraction based on pattern discovery. In *Proceedings of the 10th international conference on World Wide Web*, pages 681–688. ACM.

Huang, J., Zhou, M., and Yang, D. (2007). Extracting chatbot knowledge from online discussion forums. In *IJCAI*, volume 7, pages 423–428.

Jin, O., Liu, N. N., Zhao, K., Yu, Y., and Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 775–784. ACM.

Liu, B., Grossman, R., and Zhai, Y. (2003). Mining data records in web pages. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–606. ACM.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *EMNLP*, volume 4, pages 404–411.

Reis, D. d. C., Golgher, P. B., Silva, A. S., and Laender, A. (2004). Automatic web news extraction using tree edit distance. In *Proceedings of the 13th international conference on World Wide Web*, pages 502–511. ACM.

Surdeanu, M., Ciaramita, M., and Zaragoza, H. (2008). Learning to rank answers on large online qa collections. In *2008 Annual Meeting of the Association for Computational Linguistics*, volume 8, pages 719–727.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392.

Wang, B., Liu, B., Sun, C., Wang, X., and Sun, L. (2009). Extracting chinese question-answer pairs from online forums. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pages 1159–1164. IEEE.

Wang, J. and Wang, J. (2015). qRead: A fast and accurate article extraction method from web pages using partition features optimizations. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on*, volume 1, pages 364–371. IEEE.

Yang, J.-M., Cai, R., Wang, Y., Zhu, J., Zhang, L., and Ma, W.-Y. (2009). Incorporating site-level knowledge to extract structured data from web forums. In *Proceedings of the 18th international conference on World wide web*, pages 181–190. ACM.

Yu, H., Han, J., and Chang, K. C.-C. (2002). PEBL: Positive example based learning for web page classification using SVM. In *Proceedings of the 8th ACM SIGKDD International conference on Knowledge discovery and data mining*, pages 239–248. ACM.

Zhang, C. and Wang, J. (2017). RQAS: a rapid QA scheme with exponential elevations of keyword rankings. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2017 9th International Joint Conference on*.