# A Community Question-Answering Refinement System

Maria Soledad Pera
Computer Science Department
Brigham Young University
Provo, Utah, U.S.A.
mpera@cs.byu.edu

Yiu-Kai Ng
Computer Science Department
Brigham Young University
Provo, Utah, U.S.A.
ng@compsci.byu.edu

## ABSTRACT

Community Question Answering (CQA) websites, which archive millions of questions and answers created by CQA users to provide a rich resource of information that is missing at web search engines and QA websites, have become increasingly popular. Web users who search for answers to their questions at CQA websites, however, are often required to either (i) wait for days until other CQA users post answers to their questions which might even be incorrect, offensive, or spam, or (ii) deal with restricted answer sets created by CQA websites due to the exact-match constraint that is employed and imposed between archived questions and user-formulated questions. To automate and enhance the process of locating high-quality answers to a user's question $Q$ at a CQA website, we introduce a CQA refinement system, called $QAR$. Given $Q$, $QAR$ first retrieves a set of CQA questions $QS$ that are the same as, or similar to, $Q$ in terms of its specified information need. Thereafter, $QAR$ selects as answers to $Q$ the top-ranked answers (among the ones to the questions in $QS$) based on various similarity scores and the length of the answers. Empirical studies, which were conducted using questions provided by the Text Retrieval Conference (TREC) and Text Analysis Conference (TAC), in addition to more than four millions questions (and their corresponding answers) extracted from Yahoo! Answers, show that $QAR$ is effective in locating archived answers, if they exist, that satisfy the information need specified in $Q$. We have further assessed the performance of $QAR$ by comparing its *question-matching* and *answer-ranking* strategies with their Yahoo! Answers' counterparts and verified that $QAR$ outperforms Yahoo! Answers in (i) locating the set of questions $QS$ that have the highest degrees of similarity with $Q$ and (ii) ranking archived answers to $QS$ as answers to $Q$.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: [Information Search and Retrieval–Search process, Retrieval models]; H.3.5 Information Storage and Retrieval [Online Information Services–Web-based services]

## General Terms

Algorithm, Experimentation

## Keywords

Community question answering, word similarity measure, question matching, answer ranking

## 1. INTRODUCTION

During the last few years, web users have been turning to Community Question-Answering (CQA) websites, such as Yahoo! Answers (answers.yahoo.com), WikiAnswers (wiki.answers.com), Naver (naver.com), and AskVille (askville.amazon.com), to look for and/or provide answers to questions in diverse topics[1]. A CQA system exploits the power of human knowledge to satisfy a broad range of users' information needs and handles factoid, as well as complex[2], questions which could be very difficult, if not impossible, to be answered by conventional web search engines or existing QA systems, such as Ask.com [9]. While this repository provides a rich resource of information that is missing at popular web search engines and QA websites, locating answers to a new user's question $Q$ using question-answer pairs archived at CQA websites is a challenging task. The challenge is caused by using different wordings in formulating (the same or similar) questions by various users, which complicates the process of finding relevant archived answers to $Q$ due to the $exact\text{-}keyword\ matching$ strategy employed by existing CQA systems on archived questions (answers, respectively) and $Q$ [9, 28].

We propose to develop a CQA refinement system, denoted $QAR$, so that given a new user's question $Q$, $QAR$ identifies closely related, besides exact-matched, archived questions to $Q$ (in terms of their information needs) and chooses the highly-ranked archived answers to the identified questions as the answers to $Q$. Matched questions and their corresponding answers retrieved by $QAR$ can be extracted from any existing CQA website.

To reduce the huge number of comparisons for retrieving closely related questions to $Q$, $QAR$ identifies the most representative terms $T$, instead of using all the keywords, in $Q$ and employs a *blocking strategy* which selects and ranks archived questions that contain keywords that exactly match or are highly similar to $T$. The *similarity match* which determines the degree of resemblance

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*HT'11,* June 6–9, 2011, Eindhoven, The Netherlands.
Copyright 2011 ACM 978-1-4503-0256-2/11/06 ...$10.00.

[1] Web users have contributed millions of answers to questions at various CQA websites. As of December 2007, Yahoo! Answers collected more than 400 million answers to user-posted questions [28].

[2] Complex questions are questions inquiring opinions or advice which yield potentially multiple answers to be ranked [18].

between an archived CQA question and $Q$ is conducted using word-correlation factors. The answers $A$s to the highly-resembled archived questions are ranked using (i) the *degree of similarity* of an archived answer $A$ (in $As$) and $Q$, (ii) the *degree of similarity* of $A$ and its corresponding archived question $Q_A$, and (iii) the length of $A$. The highest-ranked answers are selected by $QAR$ to generate the set of answers to $Q$.

Unlike existing (i) question-matching methodologies [12] which identify questions that are similar to a new question $Q$, (ii) ranking strategies [25, 28] which determine the quality of answers to $Q$, and (iii) CQA systems that require users to browse through and manually choose archived answers as answers to $Q$, $QAR$ combines word-correlation factors, question-matching, and answer-ranking strategies to generate the set of ranked answers to $Q$. In addition, $QAR$ solves many of the problems currently encountered by users of existing CQA systems which include (i) waiting days for other CQA users to post answers to $Q$ and (ii) receiving no answers to $Q$.

The proposed $QAR$ fully automates the process of locating high-quality answers, if they exist, from the millions archived at a particular CQA website in response to $Q$, which minimizes its users' time and efforts involved in scanning through questions and their corresponding answers retrieved by the CQA website. We have chosen Yahoo! Answers as the source of archived questions and answers for $QAR$, since Yahoo! Answers (i) is one of the most popular CQA systems these days [18] and (ii) has established a publicly available dataset which we downloaded and used for conducting a performance evaluation on $QAR$.

The remaining of this paper is organized as follows. In Section 2, we discuss existing question-matching and answer-ranking strategies. In Section 3, we detail the design of $QAR$. In Section 4, we present the empirical studies conducted for verifying the performance of $QAR$. In Section 5, we give a concluding remark.

## 2. RELATED WORK

In this section, we discuss existing question-matching and answer-ranking approaches. The former is adopted by CQA systems to identify the most similar archived questions (in terms of their information needs) to a new user's question $Q$, whereas the latter is adopted for ranking archived answers as answers to $Q$. We compare these approaches with their $QAR$'s counterparts.

Jeon et al. [12] and Xue et al. [29] rely on trained machine-translation models to find questions that are the same or (semantically) similar to a user's question $Q$, despite their lexical mismatch. Besides using archived questions, Xue et al. [29] also consider the answers to the archived questions in performing the question-matching task. Wang et al. [27] identify questions similar to $Q$ by comparing the syntactic tree of $Q$ and its counterparts constructed using CQA questions. Cao et al. [6] introduce a question-matching framework based on the categories of questions (e.g., travel, politics, or education) archived at a CQA website. The authors first determine the category $C$ to which $Q$ belongs and rank CQA's archived questions belonged to the same category as $Q$ that are similar to $Q$. Cao et al. also search for archived questions in categories other than $C$ to which $Q$ has a high likelihood of belonging. Unlike the question-matching approaches in [6, 12, 27, 29], $QAR$ avoids any pre-processing steps, which require either training a machine-translation model, representing (CQA) questions as syntactic trees, or determining the category to which $Q$ belongs, and thus reduces the processing time spent on locating questions similar to $Q$. Besides matching $Q$ with archived questions as in [6, 12, 27, 29], $QAR$ also applies an answer-ranking strategy which ranks and selects archived answers as answers to $Q$.

In ranking CQA answers to $Q$, Suryanto et al. [25] rely on answer features, such as answer lengths and the ratings given to an answer by CQA users, along with the expertise of the users who provide answers. Bian et al. [3] take into account (i) user interaction information, such as the number of questions a user asked (answered, respectively) and the number of answers posted by a user, and (ii) community-based features, which include the positions in the ranking given to archived answers in Yahoo! Answers, to retrieve relevant answers from CQA systems. As opposed to $QAR$'s answer-ranking strategy, the ranking methodology in [3] handles factoid questions only. $QAR$ differs from the user-provided voting scheme which generates the ranking of archived answers in [25], since $QAR$'s answer-ranking strategy does not require user's involvement, is fully automated, and is semantic-driven.

While Jeon et al. [13] analyze the properties of an answer $A$, such as the length of $A$ and the number of votes $A$ receives, Agichtein et al. [1] consider the structural, textual, and community-based features of $A$, in addition to the quality[3] of its corresponding question.

In [17, 28], the authors identify the best answer to a question in a CQA system. While Lee et al. [17] develop a weighted voting scheme based on voter's credibility, which handles the plurality voting scheme problem (that is vulnerable due to random or spam voting) of CQA systems, in choosing the best answer to a question, Wang et al. [28] introduce a method based on analogical reasoning that uses (i) a set of user-provided question-answer pairs in which high-quality, incorrect, and spam answers have been previously identified and (ii) a Bayesian logistic regression model to determine a score for each candidate answer to a question. Unlike $QAR$'s ranking strategy, the methodologies proposed in [17, 28] are based on user-feedback information, which may not always be available.

The existing approaches discussed in this section either locate similar questions with respect to a given user's question or rank answers retrieved by CQA systems, but not both. $QAR$, on the other hand, is (to the best of our knowledge) the only approach that combines these two tasks into a single process.

## 3. $QAR$, OUR PROPOSED QA REFINEMENT SYSTEM

In this section, we introduce $QAR$ which matches questions archived in a CQA system with a new user's question $Q$ and extracts and ranks answers to the matched questions as answers to $Q$. We first define the word-correlation factors which indicate the degree of similarity of any two question/answer keywords, the measures that $QAR$ uses for matching questions and ranking answers (in Section 3.1). Thereafter, we discuss the question-matching strategy (in Section 3.2) employed by $QAR$, and introduce $QAR$'s answer-ranking strategy (in Section 3.3).

### 3.1 Word-Correlation Factors

$QAR$ relies on the pre-computed word-correlation factors in the word-correlation matrix [16] for matching archived questions with, and ranking answers to, $Q$. The word-correlation factors were generated using a set of approximately 880,000 Wikipedia documents (http://wikipedia.org), and each correlation factor indicates the *degree of similarity* of the two corresponding words[4] based on their

[3] In defining the quality of a question, Agichtein et al. [1] consider a variety of semantic features, which include correct use of punctuation, misspellings, and grammatical properties, to name a few.

[4] Words in the Wikipedia documents were *stemmed* [9] (i.e., reduced to their grammatical roots) after all the stopwords [9], such as articles, conjunctions, and prepositions, which do not play a sig-

(i) *frequency of co-occurrence* and (ii) *relative distances* in each Wikipedia document.

Wikipedia documents were chosen for constructing the word-correlation matrix, since they were written by more than 89,000 authors (i) with different writing styles, (ii) using various terminologies that cover a wide range of topics, and (iii) with diverse word usage and content. Furthermore, the words in the matrix are common words in the English language that appear in various online English dictionaries, such as 12dicts-4.0 (prdownloads.source forge.net/wordlist/12dicts-4.0.zip), Ispell (cs.ucla.edu/geoff/ispell. html), and BigDict (packetstormsecurity.nl/Crackers/bigdict.gz).

The word-correlation matrix is a 57,908 × 57,908 symmetric matrix, since the word-correlation factors $wcf(i, j)$ and $wcf(j, i)$ are equal, where $i$ and $j$ are any two given words, and $wcf(i, j)$ reflects how closely related $i$ and $j$ are, and is defined as

$$wcf(i,j) = \frac{\sum_{D \in Wiki} \sum_{w_i \in D} \sum_{w_j \in D} \frac{1}{d(w_i,w_j)+1}}{N_i \times N_j} \qquad (1)$$

where $Wiki$ is the set of Wikipedia documents, $w_i$ ($w_j$, respectively) is an occurrence of the word $i$ ($j$, respectively) in a Wikipedia document $D$, $d(w_i, w_j)$ is the *distance*, i.e., the number of words, between $w_i$ and $w_j$ in $D$ such that $d(w_i, w_j) = \infty$, if either $w_i \notin D$ or $w_j \notin D$, and $N_i$ ($N_i$, respectively) is the number of times word $i$ (word $j$, respectively) appeared in $Wiki$.

Compared with synonyms and related words compiled by WordNet (wordnet.princeton.edu) in which pairs of words are not assigned similarity weights, word-correlation factors provide a more sophisticated measure of word similarity. (For an in-depth discussion on the word-correlation factors and a comparison with alternative correlation measures for determining word-similarity, see [16].)

Note that in identifying archived questions that are similar to a new user's question (as discussed in Section 3.2.2), $QAR$ adopts a *reduced* version of the word-correlation matrix. The reduced word-correlation matrix contains 13% of the most frequently-occurring word pairs (based on their frequencies of occurrence in the Wikipedia documents), which was empirically determined as discussed in [10], and for the remaining 87% of the less-frequently-occurring word pairs only exact-matched word-correlation factor, i.e., 1, is used. The distribution of the word-correlation factors among different word pairs in the reduced matrix is illustrated in Figure 1, which shows that the word-correlation factors that are not exact matches are in the range of $1 \times 10^{-4}$ and $1 \times 10^{-6}$, and word pairs with a word-correlation factor no less than $1 \times 10^{-4}$ are treated as highly similar, whereas word pairs with lower word-correlation factors are treated as less similar.

## 3.2 The Question-Matching Strategy

In this section, we discuss the various steps invoked by $QAR$ in matching archived questions with a new user's question.

### 3.2.1 User's Question Representation

Since users' questions tend to be lengthy [2], $QAR$ adopts the features proposed by Bendersky et al. [2] for extracting the *most representative* keywords in verbose natural language questions to capture the information needs specified in a question, which in turn has a positive impact on the retrieval performance associated with verbose queries, i.e., retrieving archived questions similar to a user's question in our case. $QAR$ weights each keyword $K$ in a question $Q$ using the features defined in [2]: (i) $K$ is *capitalized* (with the weight of "1") or not (with the weight of "0"), since a capitalized keyword is assumed to capture important information in $Q$, (ii) the *frequency of occurrence* of $K$ in a corpus $C$, since frequently occurred terms are assumed to be more representative of the content of $Q$ than less-frequent ones, (iii) the *inverted document frequency* (IDF) of $K$ in $C$, since IDF is commonly used in information retrieval as a weighting function [23], (iv) the *residual inverted document frequency* of $K$ in $C$, which is the difference between the observed IDF and the value predicted by a Poisson model [8] of $K$, since the difference reflects the degree of significance of $K$ in representing the information needs expressed in $Q$, (v) the *weighted information gain* (WIG) of $K$, since WIG measures the change in information on the quality of the retrieval in response to $K$ [31] and serves as an indicator of representative keywords, (vi) the *frequency of occurrence* of $K$ based on the Google unigram counts [5], which can be a more reliable frequency estimator than the frequency of $K$ in $C$, since the latter vary depending on the size of $C$, (vii) the *number* of times $K$ appears as part of a question in $C$, and (viii) the *number* of times $K$ is an exact question in $C$.

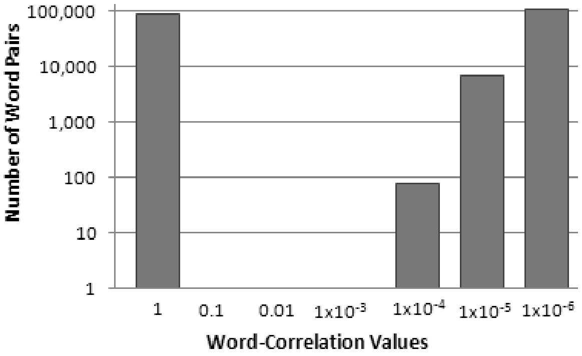


**Figure 1: Distribution of the word-correlation factors in the reduced word-correlation matrix**

To compute a score, i.e., $Rank(K)$, for each keyword $K$ in $Q$, which identifies the *degree of significance* of $K$ in representing the information need specified in $Q$, $QAR$ uses the *Odds ratio* [14] ($Odds$ for short), which is defined as the ratio of the probability ($p$) that an event occurs to the probability (1 - $p$) that it does not.

$$Odds(H) = \frac{p(H)}{1 - p(H)} \qquad (2)$$

where $Odds(H)$ measures the predictive or prospective support according to a hypothesis $H$ by the prior knowledge $p(H)$ alone to determine the strength of a belief, which is based on the feature values listed above in our case.

In computing $Rank(K)$ of a keyword $K$ in $Q$, $QAR$ relies on the product of the feature values computed for $K$ in $Q$, i.e., $p(H)$ in Equation 2, which determines the significance of $K$ in $Q$. Since the feature values are in different numerical ranges, $QAR$ normalizes the feature values so that each score is bounded between 0 and 1, and they are weighted equally. $Rank(K)$ is defined as

$$Rank(K) = \frac{\prod_{i=1}^{8} \frac{Feature_i(K)}{argmax_i Feature_i(K)}}{1 - \prod_{i=1}^{8} \frac{Feature_i(K)}{argmax_i Feature_i(K)}} \qquad (3)$$

where $Feature_i(K)$ is the $i^{th}$ ($1 \leq$ i $\leq$ 8) feature score for $K$ in

nificant role in representing the content of a document, were removed. As a side-effect, the stopword removal and stemming process significantly reduces the number of (key)words to be considered. From now on, unless stated otherwise, (key)words refer to *non-stop*, *stemmed words*.

$Q$ and $argmax_i Feature_i(K)$ is a function that identifies the $i^{th}$ feature score for $K$ in $Q$ with the highest score which is the normalization factor that bounds the feature scores between 0 and 1.

Having computed $Rank(K)$ and based on the analysis that an average query includes 2.6 terms [24], the *top*-3 highest-ranked keywords of $Q$ are chosen for representing $Q$ in selecting CQA questions (see details in Section 3.2.2). Processing the top-3 highly-ranked keywords, instead of all the (non-)stopwords in $Q$, significantly reduces the question evaluation time of $Q$.

### 3.2.2 Selecting Similar Archived Questions

As previously stated, a CQA system archives millions of questions and thus it is not practical to compare each question in the system with a new user's question $Q$ to find archived questions that match $Q$. To avoid computing the *degrees of resemblance* between $Q$ and each of the archived CQA questions so that question processing time can be further minimized, $QAR$ chooses a subset $S$ of archived questions, if they exist, that have a high degree of similarity to $Q$. In accomplishing this task, $QAR$ applies a blocking strategy[5] to retrieve CQA questions that include keywords that either *exactly match* or are *highly similar* to each of the top-3 representative keywords in $Q$. In other words, to include an archived question $Q$' in $S$ (to yield the subset of questions highly-likely relevant to $Q$), each of the top-3 keywords $k$ representing $Q$ either (i) exactly matches a keyword in $Q$' or (ii) the correlation factor of a keyword in $Q$' and $k$ is in the reduced word-correlation matrix (as defined in Section 3.1).

Pera et al. [21] have verified that by using the reduced word-correlation matrix, as opposed to the word-correlation matrix introduced in Section 3.1, it is possible to select a subset of items, i.e., questions in our case, to be evaluated and decrease the item-matching processing time without affecting the matching accuracy.

### 3.2.3 Ordering Matched Questions

Having determined the subset $S$ of CQA questions that are similar to $Q$, $QAR$ ranks the questions in $S$ to identify the ones with the highest degree of resemblance to $Q$. The *degree of resemblance* between $Q$ and each question $Q^{'}$ in $S$ is computed as follows:

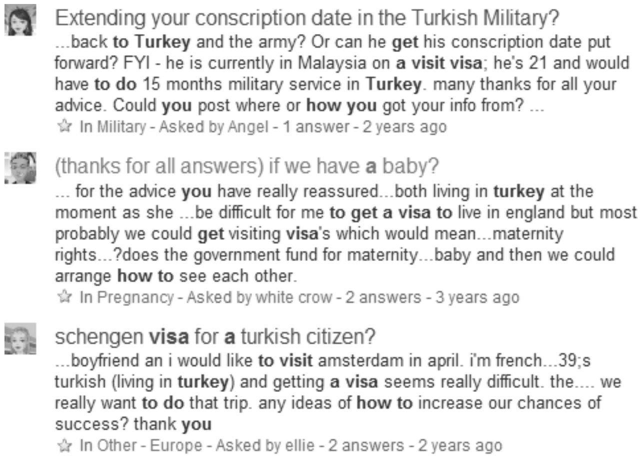
$$Sim(Q, Q^{'}) = \sum_{i=1}^{n} \sum_{j=1}^{m} wcf(q_i, q_j^{'}) \quad (4)$$

where $n$ ($m$, respectively) is the number of keywords in $Q$ ($Q^{'}$, respectively), $q_i$ ($q_j^{'}$, respectively) is a keyword in $Q$ ($Q^{'}$, respectively), and $wcf(q_i, q_j^{'})$ is the word-correlation factor of $q_i$ and $q_j^{'}$, as defined in Equation 1. Note that all the keywords in $Q$, not just the top-3 most representative ones in $Q$ which are simply used to identify candidate questions, i.e., questions similar to $Q$, are considered in computing $Sim(Q, Q^{'})$, which should yield a more reliable similarity measure compared with using only the top-3 keywords in questions of $S$, a relatively small subset.

The length of $Q^{'}$ can potentially affect $Sim(Q, Q^{'})$, since the $longer$ $Q^{'}$ is, the $higher$ the $Sim(Q, Q^{'})$ value could be, which could create a *bias* in its degree of resemblance to $Q$. $QAR$ normalizes $Sim(Q, Q^{'})$ as follows:

$$NSim(Q, Q^{'}) = \frac{Sim(Q, Q^{'})}{m} \quad (5)$$

where $m$ and $Sim(Q, Q^{'})$ are as defined in Equation 4.

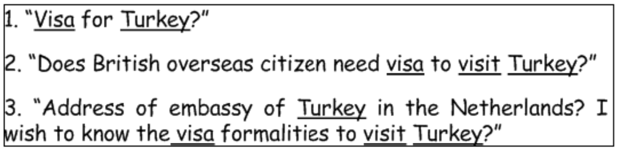


Extending your conscription date in the Turkish Military?
...back **to Turkey** and the army? Or can he **get** his conscription date put forward? FYI - he is currently in Malaysia on **a visit visa**; he's 21 and would have **to do** 15 months military service in **Turkey**. many thanks for all your advice. Could **you** post where or **how you** got your info from? ...
☆ In Military - Asked by Angel - 1 answer - 2 years ago

(thanks for all answers) if we have **a** baby?
... for the advice **you** have really reassured...both living in **turkey** at the moment as she ...be difficult for me **to get a visa to** live in england but most probably we could **get** visiting **visa**'s which would mean...maternity rights...?does the government fund for maternity...baby and then we could arrange **how to** see each other.
☆ In Pregnancy - Asked by white crow - 2 answers - 3 years ago

schengen **visa** for **a** turkish citizen?
...boyfriend an i would like **to visit** amsterdam in april. i'm french...39;s turkish (living in **turkey**) and getting **a visa** seems really difficult. the.... we really want **to do** that trip. any ideas of **how to** increase our chances of success? thank **you**
☆ In Other - Europe - Asked by ellie - 2 answers - 2 years ago

**Figure 2: Top-3 questions (out of the top-10) retrieved by Yahoo! Answers in response to the question $Q$, "How do you get a visa to visit Turkey?"**



1. "Visa for Turkey?"

2. "Does British overseas citizen need visa to visit Turkey?"

3. "Address of embassy of Turkey in the Netherlands? I wish to know the visa formalities to visit Turkey?"

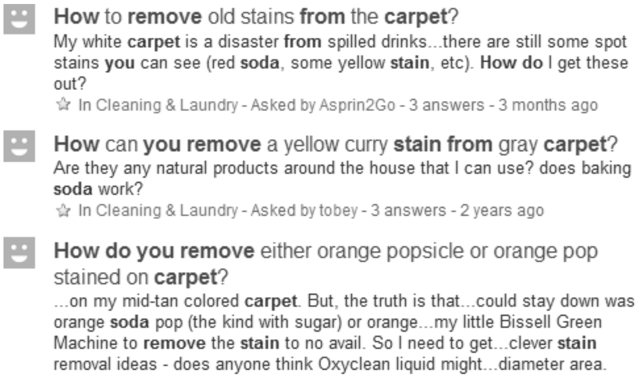
**Figure 3: Top-3 questions (among the top-10) retrieved by $QAR$ for the question $Q$, "How do you get a visa to visit Turkey?", in which words that exactly match the top-3 words, 'visa", "visit", and "Turkey", representing $Q$ are underlined**

Since web search engine users often view only the first 10 retrieved results when performing a search [11], we only consider up to the top-10 CQA questions (in $S$) with the highest $NSim$ values as the most (semantically) similar archived questions to $Q$.

EXAMPLE 1. Consider the question $Q$, "How do you get a visa to visit Turkey?". Both Yahoo! Answers and $QAR$ identify (related) archived questions for $Q$. While the first ranked question retrieved by Yahoo! Answers (on June 30, 2010), as shown in Figure 2, is "Extending your conscription date in the Turkish Military?", which does not match the information needs specified in $Q$, $QAR$ extracts the first ranked question (as shown in Figure 3), "Visa for Turkey?", which does. Moreover, the $2^{nd}$ and $3^{rd}$ questions retrieved by Yahoo! Answers for $Q$ (as shown in Figure 2) do not match $Q$, since the questions were posted by users living in Turkey who were interested in applying for visas to other countries. The $2^{nd}$ and $3^{rd}$ questions retrieved by $QAR$ (as shown in Figure 3), however, are related to $Q$, since they both inquire information on applying for visas to visit Turkey. □

EXAMPLE 2. Consider another question $Q_E$, "How do you remove soda stain from carpet?". Figure 4 (Figure 5, respectively) shows the top-3 (among the top-10) questions retrieved and ranked by Yahoo! Answers ($QAR$, respectively) in response to $Q_E$. Unlike the top-3 questions retrieved by $QAR$, which are highly similar to $Q_E$, since they match the same information need as specified in $Q_E$, the questions retrieved by Yahoo! Answers inquire on how to remove from a carpet either general stains or curry stains, rather than soda stains, and thus are not (closely) related to the original question $Q_E$. □

[5] A *blocking strategy* [15] is a filtering technique which reduces the potentially very large number of records to be compared [7].

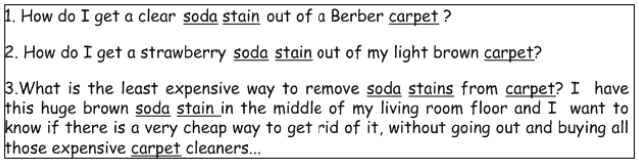


**How** to **remove** old stains **from** the **carpet**?
My white **carpet** is a disaster **from** spilled drinks...there are still some spot stains **you** can see (red **soda**, some yellow **stain**, etc). **How do** I get these out?
☆ In Cleaning & Laundry - Asked by Asprin2Go - 3 answers - 3 months ago

**How** can **you remove** a yellow curry **stain from** gray **carpet**?
Are they any natural products around the house that I can use? does baking **soda** work?
☆ In Cleaning & Laundry - Asked by tobey - 3 answers - 2 years ago

**How do you remove** either orange popsicle or orange pop stained on **carpet**?
...on my mid-tan colored **carpet**. But, the truth is that...could stay down was orange **soda** pop (the kind with sugar) or orange...my little Bissell Green Machine to **remove** the **stain** to no avail. So I need to get...clever **stain** removal ideas - does anyone think Oxyclean liquid might...diameter area.

**Figure 4: Top-3 questions identified by Yahoo! Answers in response to the question $Q_E$, "How do you remove soda stain from carpet?"**



1. How do I get a clear soda stain out of a Berber carpet ?

2. How do I get a strawberry soda stain out of my light brown carpet?

3.What is the least expensive way to remove soda stains from carpet? I have this huge brown soda stain in the middle of my living room floor and I want to know if there is a very cheap way to get rid of it, without going out and buying all those expensive carpet cleaners...

**Figure 5: Top-3 questions identified by $QAR$ in response to $Q_E$, "How do you remove soda stain from carpet?", in which keywords that exactly match the top-3 most representative words, i.e., "soda", "stain", and "carpet", in $Q_E$ are underlined**

Notice that in determining the similarity among questions, $QAR$ accumulates word-correlation factors, instead of depending on traditional document similarity measures (e.g., cosine similarity), since the latter have been shown to perform poorly in handling short texts, which include very few, if any, overlapping terms [22].

## 3.3 The Answer-Ranking Strategy

Having determined the set $S$ of the top-10 archived questions most similar to a new user's question $Q$, $QAR$ proceeds to $rank$ each archived answer $A$ to each question in $S$ to determine its relative degree of satisfaction in answering (the information needs specified in) $Q$. To determine the likelihood of $A$ in answering $Q$, $QAR$ considers (i) the similarity between $A$ and $Q$, i.e., $NSim(A, Q)$, (ii) the similarity between $A$ and its corresponding question $Q_A$, a question in $S$, i.e., $NSim(A, Q_A)$, and (iii) the length of $A$, denoted $Length(A)$.

As claimed by Tu et al. [26], one of the major challenges in identifying correct answers $As$ in response to a user's question $Q$ is the lexical gap between $Q$ and $As$, which is caused by the *textual mismatch* between $Q$ and $As$, i.e., $Q$ includes words that do not necessarily occur in $As$. $QAR$ relies on the $word\text{-}correlation\ factors$ (introduced in Section 3.1) to determine the similarity between questions and answers, which relaxes the *exact-keyword matching* constraint imposed by CQA systems in locating answers that respond to the information needs specified in a particular question. $NSim(A, Q)$ ($NSim(A, Q_A)$, respectively), as defined in Equation 5 in which the keywords in $Q$ and $Q'$ are the keywords in $A$ and $Q$ ($Q_A$, respectively), indicates to what degree $A$ satisfies the information needs specified in $Q$ ($Q_A$, respectively) and is based on the word-correlation factors of each keyword in $A$ with respect to each keyword in $Q$ ($Q_A$, respectively).

$QAR$ relies on $NSim(A, Q)$, since the highest the similarity score between $A$ and $Q$, the more likely $A$ is an archived answer that satisfies the information needs specified in $Q$. $NSim(A, Q_A)$, on the other hand, reflects the degree of confidence of $A$ in answering $Q_A$. $QAR$ computes the similarity between $A$ and its corresponding question, as opposed to using the actual ranking determined by the ratio of positive and negative votes given to $A$ by CQA users, since as stated in [4, 28], while users' votes can provide indicators of the quality and readability of an answer, they are not always reliable due to the existence of bad or fraudulent votes, i.e., spam votes. More importantly, Suryanto et al. [25] have verified that measuring the relevance of an answer using its question is a better alternative than considering answer attributes, such as the number of times an answer is recommended by other users or the number of times a user prints or copies an answer.

$QAR$ also employs $Length(A)$, which returns the number of keywords in $A$, as a factor in ranking $A$ with respect to $Q$, since as stated and verified in [13], good, i.e., high-quality, answers are usually longer than bad answers, i.e., spam answers in CQA systems, which include answers such as "I don't know" or "Nothing new".

$QAR$ computes a ranking score for $A$, denoted $RankAns(A)$, which reflects the relative degree of satisfaction of $A$ in providing the information needs expressed in $Q$. The ranking score is calculated by combining (the scores of) each of the measures[6] previously described using the *Stanford Certainty Factor* (SCF) [19], which is a measure that integrates different assessments, i.e., various answer scores in our case, to determine the $strength$ of a hypothesis, i.e., the effectiveness of $A$ in answering $Q$ in our case. The formal definition of SCF is given as follows:

$$SCF(C) = \frac{SCF(R_1) + SCF(R_2)}{1 - Min\{SCF(R_1), SCF(R_2)\}} \qquad (6)$$

where $R_1$ and $R_2$ are two hypotheses that reach the same conclusion $C$, and SCF is the Stanford certainty factor (i.e., confidence measure) of $C$, which is a monotonically increasing (decreasing) function of combined assumptions for computing the confidence measure of $C$.

Using the SCF equation, $QAR$ combines the various measures related to $A$ to yield the overall $RankAns$ score of $A$ as follows:

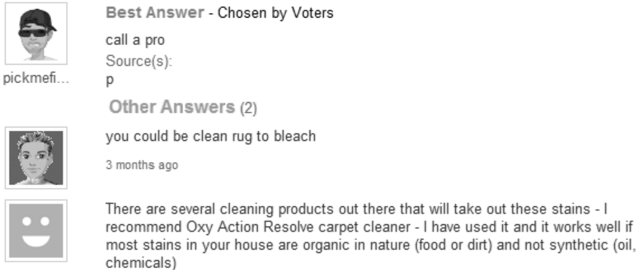
$$RankAns(A) = \frac{NSim(A,Q) + NSim(A,Q_A) + Length(A)}{1 - Min\{NSim(A,Q), NSim(A,Q_A), Length(A)\}} \qquad (7)$$

Since, as previously stated, it is a common practice for web users to view only the top-10 retrieved results when performing a search [11], $QAR$ displays up to the top-10 retrieved answers with the highest $RankAns$ scores as the answers to $Q$.

EXAMPLE 3. The top-3 (out of the top-10) ranked answers retrieved by Yahoo! Answers[7] ($QAR$, respectively) as answers to $Q_E$ in Example 2 are shown in Figure 6 (Figure 7, respectively). It is clear that the top-3 answers to $Q_E$ retrieved by $QAR$ satisfy the information need specified in $Q_E$, as opposed to the answers chosen by Yahoo! Answers, in which only *one* out of the top-3 answers, i.e., Answer 3 in Figure 6, provides an answer that could

[6] Since the measures employed in computing the ranking score of a given answer are in different numerical ranges, $QAR$ scales all the measures using a $log_{10}$ function so that they are in the same range.

[7] In identifying the relative order of answers retrieved by Yahoo! Answers for illustration purposes, we simulate an "intelligent" user who given a question $Q$ always selects the most relevant question with respect to $Q$ retrieved by Yahoo! Answers, a common practice. See Section 4.2.3 for details.



Best Answer - Chosen by Voters

call a pro

Source(s):

p

pickmefi...

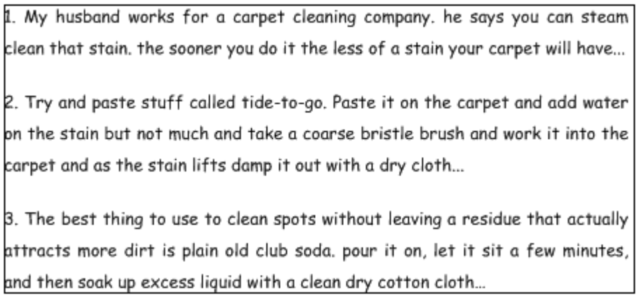
Other Answers (2)

you could be clean rug to bleach

3 months ago

There are several cleaning products out there that will take out these stains - I recommend Oxy Action Resolve carpet cleaner - I have used it and it works well if most stains in your house are organic in nature (food or dirt) and not synthetic (oil, chemicals)

**Figure 6: The top-3 ranked answers selected by Yahoo! Answers, which are answers to the archived questions in Figure 4**



1. My husband works for a carpet cleaning company. he says you can steam clean that stain. the sooner you do it the less of a stain your carpet will have...

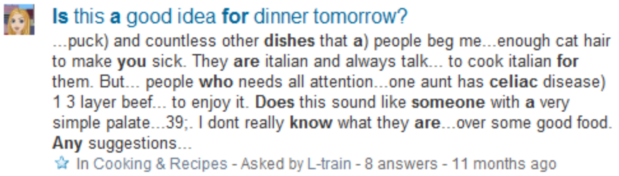
2. Try and paste stuff called tide-to-go. Paste it on the carpet and add water on the stain but not much and take a coarse bristle brush and work it into the carpet and as the stain lifts damp it out with a dry cloth...

3. The best thing to use to clean spots without leaving a residue that actually attracts more dirt is plain old club soda. pour it on, let it sit a few minutes, and then soak up excess liquid with a clean dry cotton cloth...

**Figure 7: The top-3 ranked answers retrieved by $QAR$, which are answers to the archived questions in Figure 5**

be considered adequate for $Q_E$. More importantly, while the three answers shown in Figure 7 offer different suggestions for removing *soda stain* from *carpets* (as specified in $Q_E$), the third answer extracted by Yahoo! Answers (as shown in Figure 6) discusses how to remove general stains from carpets, which is less informative and useful than the answers retrieved by $QAR$. □

EXAMPLE 4. Consider the question $Q_I$, "Do you know any dishes for someone who is a celiac?". As shown in Figure 8, Yahoo! Answers retrieves a single archived question in response to $Q_I$, which does not match the information need specified in $Q_I$, since the retrieved question inquiries on suggested dishes to serve for a dinner party. $QAR$, on the other hand, identifies archived questions which match the same information need as specified in $Q_I$, including keywords similar to "dishes", i.e., "ingredients" and "recipes", and exactly-matched keyword "celiac" in $Q_I$ (see the top-3 archived questions shown in Figure 9). In fact, each of the top-10 answers retrieved by $QAR$ in response to $Q_I$ is relevant to $Q_I$, which include suggested recipes for those who suffer from the celiac disease. The top-3 answers retrieved by $QAR$ in response to $Q_I$ are shown in Figure 10. □

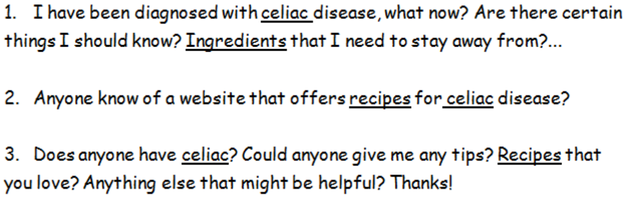


**Is** this **a** good idea **for** dinner tomorrow?

...puck) and countless other **dishes** that **a**) people beg me...enough cat hair to make **you** sick. They **are** italian and always talk... to cook italian **for** them. But... people **who** needs all attention...one aunt has **celiac** disease) 1 3 layer beef... to enjoy it. **Does** this sound like **someone** with **a** very simple palate...39;. I dont really **know** what they **are**...over some good food. **Any** suggestions...

☆ In Cooking & Recipes - Asked by L-train - 8 answers - 11 months ago

**Figure 8: An archived question retrieved by Yahoo! Answers in response to question $Q_I$, "Do you know any dishes for someone who is a celiac?"**



1. I have been diagnosed with celiac disease, what now? Are there certain things I should know? Ingredients that I need to stay away from?...

2. Anyone know of a website that offers recipes for celiac disease?

3. Does anyone have celiac? Could anyone give me any tips? Recipes that you love? Anything else that might be helpful? Thanks!

**Figure 9: Top-3 archived questions retrieved by $QAR$ in response to question $Q_I$, "Do you know any dishes for someone who is a celiac?", in which keywords that exactly-match or are highly similar to the keywords representing $Q_I$, i.e., "celiac" and "dishes", are underlined**

# 4. EXPERIMENTAL RESULTS

In this section, we first introduce the datasets and metrics used for assessing the performance of $QAR$ (in Sections 4.1 and 4.2, respectively). Thereafter, we detail the empirical studies conducted for verifying the effectiveness of $QAR$ in (i) matching archived questions with a new user's question $Q$ (in Section 4.3.1) and (ii) retrieving and ranking archived CQA answers which serve as answers to $Q$ (in Section 4.3.2).

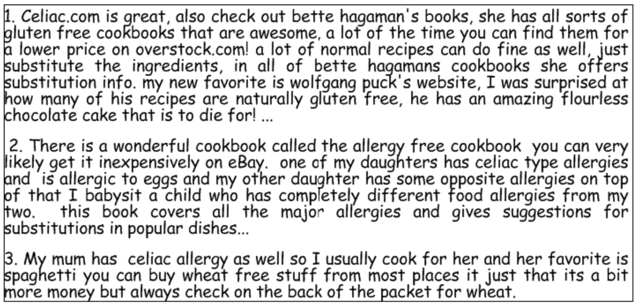
## 4.1 Datasets

As mentioned in the introduction, we consider the Yahoo! Answers Comprehensive Questions and Answers dataset [30], denoted $YA\text{-}Data$, as the *source* of questions and answers used by $QAR$ for *matching* a new user's question $Q$ and *ranking* (retrieved) archived answers as answers to $Q$. $YA\text{-}Data$ consists of 4,483,032 questions (and their corresponding answers) collected by Yahoo! Answers as of October 2007. In addition to each question $Q$ and its answers, the dataset contains metadata of $Q$ which indicates the best answer to $Q$.

Besides $YA\text{-}Data$, we have followed the evaluation strategy presented in [3] by considering another set of 300 questions, denoted $QA\text{-}dataset$. The questions in $QA\text{-}dataset$ play the role of *new users' questions* for objectively evaluating the effectiveness of the *question-matching* and *answer-ranking* strategies of $QAR$ and Yahoo! Answers (for comparison purpose), respectively. $QA\text{-}dataset$ consists of questions provided by the 2004 Text Retrieval Conference, TREC (http://trec.nist.gov/data/qa/t2004_qadata.html), in addition to a subset of "squishy", i.e., opinion, questions provided by the Opinion QA task of the 2008 Text Analysis Conference, TAC (http://www.nist.gov/tac/data/index.html). The (squishy) questions provided by TAC refer to various topics covered in the Blog06 document collection, i.e., a collection of blog posts downloaded from the Web between December 2005 and February 2006. Since Yahoo! Answers does not address all the topics covered in Blog06, we included, as part of $QA\text{-}dataset$, the TAC questions for which their corresponding topics are covered in Yahoo! Answers. During the performance evaluation process, while archived questions and answers retrieved by $QAR$ in response to each question in $QA\text{-}dataset$ came from $YA\text{-}Data$, the corresponding sets of questions and answers retrieved by Yahoo! Answers were extracted directly from the current Yahoo! Answers website with questions and answers archived up till January 11, 2011.

## 4.2 Evaluation Metrics

To evaluate the performance of $QAR$ (Yahoo! Answers, respectively) in matching questions and ranking answers, we rely on well-known information retrieval measures that include *Accuracy*, *Mean Reciprocal Rank*, *Precision at K*, and *Mean Average Preci-*



1. Celiac.com is great, also check out bette hagaman's books, she has all sorts of gluten free cookbooks that are awesome, a lot of the time you can find them for a lower price on overstock.com! a lot of normal recipes can do fine as well, just substitute the ingredients, in all of bette hagamans cookbooks she offers substitution info. my new favorite is wolfgang puck's website, I was surprised at how many of his recipes are naturally gluten free, he has an amazing flourless chocolate cake that is to die for! ...

2. There is a wonderful cookbook called the allergy free cookbook  you can very likely get it inexpensively on eBay.  one of my daughters has celiac type allergies and  is allergic to eggs and my other daughter has some opposite allergies on top of that I babysit a child who has completely different food allergies from my two.   this book covers all the major allergies and gives suggestions for substitutions in popular dishes...

3. My mum has  celiac allergy as well so I usually cook for her and her favorite is spaghetti you can buy wheat free stuff from most places it just that its a bit more money but always check on the back of the packet for wheat.

**Figure 10: Top-3 archived answers retrieved by $QAR$, which are answers to the archived question in Figure 9**

*sion*, which are commonly used for assessing the performance of a CQA system [3, 28, 29].

### 4.2.1 Accuracy on Question Matching

To assess the effectiveness of $QAR$ (Yahoo! Answers, respectively) in identifying archived CQA questions that are the same as (or closely related to) a new user's question $Q$, we evaluate the accuracy of the question-matching strategy of $QAR$ (Yahoo! Answers, respectively) using the *accuracy ratio* defined below.

$$Accuracy = \frac{Number\_of\_Related\_Questions}{Number\_of\_Retrieved\_Questions} \tag{8}$$

where *Number_of_Retrieved_Questions* is the number of questions retrieved by $QAR$ (Yahoo! Answers, respectively)[8] and $Number\_of\_Related\_Questions$ is the number of questions that are relevant[9] to $Q$.

### 4.2.2 Assessing the Answer Ranking Strategy

To determine the *relevance* of the answers retrieved (either by $QAR$ or Yahoo! Answers) for each TREC question in $QA$-$dataset$ we rely on the $answer\ patterns$ (http://trec.nist.gov/data/qa/2004_qadata/04.patterns.zip) provided by TREC, which are also used in [3]. The answer pattern $P$ of a TREC question $Q$, which is a $sequence\ of\ phrases$, is compared against each of the answers $A$ retrieved by either $QAR$ or provided by Yahoo! Answers in response to $Q$. If there is a (string) match between any phrase in $P$ and the keywords in $A$, then $A$ is labeled as *relevant* to $Q$; otherwise, $A$ is labeled as *non-relevant* to $Q$. The tens of questions in $QA$-$dataset$ provided by TAC are opinion questions that are subjective by nature and thus their relevance cannot be determined the same as TREC in which answer patterns are provided. As a result, we rely on *independent appraisers* to determine the (non-)relevance of each answer retrieved by $QAR$ (Yahoo! Answers, respectively) in response to its corresponding TAC question in $QA$-$dataset$.

To measure the $ranking$ of archived CQA *answers* extracted by $QAR$ (Yahoo! Answers, respectively), which have been identified as *relevant* answers to a new user's question by $QAR$ (Yahoo! Answers, respectively) earlier, we rely on well-known ranking measures as defined below.

**Mean Reciprocal Rank (MRR)**

The $MRR$ of the ranked answers retrieved (by either $QAR$ or Yahoo! Answers) is the averaged sum of the ranking values for each (new user's) question $Q$ such that each ranking value is either the reciprocal of the ranking position of the *first relevant* answer among the top-10 retrieved answers to $Q$, if it exists, or 0, otherwise.

$$MRR = \frac{1}{|Q_r|} \sum_{q \in Q_r} \frac{1}{r_q} \tag{9}$$

where $Q_r$ is the set of questions in $QA$-$dataset$, $|Q_r|$ is the total number of questions in $Q_r$, $q$ is one of the questions in $Q_r$, and $r_q$ is the (position in the) rank of the *first relevant* answer to $q$, if it exists.

**Precision at K (P@K)**

The $P@K$ value [20] quantifies the top-$K$ ranked answers to a (new user's) question $Q$ in terms of their relevance with respect to $Q$, which measures the overall user's satisfaction with the top-$K$ results.

$$P@K = \frac{1}{|Q_r|} \sum_{q \in Q_r} \frac{|R_q|}{K} \tag{10}$$

where $K$ (= 1, 5, and 10 in our study, which evaluate the relevance of the answers retrieved at the $top$, in the $middle$, and $overall$ in the ranking, respectively) is the (pre-defined) number of retrieved answers to be considered, $Q_r$, $|Q_r|$ and $q$ are as defined in Equation 9, and $|R_q|$ ($1 \leq |R_q| \leq K$) is the number of top-$K$ retrieved answers that are relevant to $q$.

**Mean Average Precision (MAP)**

The $MAP$ metric evaluates the (i) *average precision* of the retrieved answers and (ii) *effectiveness* of the *ranking* approach adopted by $QAR$ (Yahoo! Answers, respectively), which should position higher in the ranking the answers with higher degree of relevance to the corresponding question. $MAP$ is defined as

$$MAP = \frac{1}{|Q_r|} \times \sum_{q \in Q_r} \frac{\sum_{r=1}^{N} P@r \times rel(r)}{|R_q|} \tag{11}$$

where $|Q_r|$ and $q$ are as defined in Equation 9, $|R_q|$ ($1 \leq |R_q| \leq 10$) is as defined in Equation 10, $N$ ($1 \leq N \leq 10$) is the number of answers retrieved for $q$, $r$ is a position in the ranking (from 1 up till 10, the largest possible value), $rel(r)$ is a binary function of '1' or '0', which indicates the relevance or non-relevance of the $r^{th}$ ranked answer, respectively, and $P@r$ is the *precision* (as defined in Equation 10 without restricting $K$ being 1, 5, or 10 only) at the given cut-off rank $r$.

The ideal value of $MAP$ is 1, which indicates that all the retrieved answers are relevant to its corresponding question, and the closer $MAP$ is to 1, the better the retrieval and ranking performance of the corresponding (CQA) system is.

[8] As stated in Section 3.2.3, we consider only (up to) the top-10 most similar (or same) questions retrieved with respect to a user's question, if they are available.

[9] To the best of our knowledge, there is no dataset which can serve as benchmark data that identify the relevance of a set of questions with respect to another one. We rely on *independent appraisers* who manually examined each of the top-10 questions $Q^{'}$ retrieved by Yahoo! Answers ($QAR$, respectively) for each $QA$-$dataset$ question $Q$ to determine the relevance of $Q^{'}$ with respect to $Q$.

### 4.2.3 Baseline Evaluation Metrics Using Yahoo! Answers

We evaluate the quality of the answers retrieved by $QAR$ and compare their results with the ones retrieved by Yahoo! Answers. Yahoo! Answers relies on the votes assigned to archived answers casted by Yahoo! Answers users such that the *best* answer to an archived question $Q^{'}$ is positioned at the *top* of the answer list and the *subsequent* answers to $Q^{'}$ are ranked in decreasing order according to the number of votes they received. Since in response for each $QA\text{-}dataset$ question $Q$, Yahoo! Answers provides a list of archived questions with respect to $Q$, denoted $YA\text{-}Qs$, i.e., $Y_{Q_a}$, $Y_{Q_b}$, …, $Y_{Q_x}$, and their corresponding answers, i.e., $Y^1_{Q_a}$, …, $Y^n_{Q_a}$, $Y^1_{Q_b}$, …, $Y^m_{Q_b}$, …, $Y^1_{Q_x}$, …, $Y^z_{Q_x}$, we consider multiple alternatives for calculating $MRR$, $P@K$, and $MAP$ values of the answers retrieved by Yahoo! Answers, as suggested in [3], which are defined below.

**MRR-MAX**

In applying Equation 9 to compute the MRR score of Yahoo! Answers using the $MAX$ method, $r_q$ of question $q$ in $QA\text{-}dataset$ is the $highest\ ranking\ position$ among the ranking positions of the *first relevant answer* to each question $Y_{Q_a}$, $Y_{Q_b}$, …, $Y_{Q_x}$ in $YA\text{-}Qs$ of $q$. This baseline simulates an "intelligent" user who always selects the highest-ranked relevant answer to the most relevant question (in $YA\text{-}Qs$) retrieved by Yahoo! Answers.

**MRR-STRICT**

Using Equation 9 to compute the MRR score of Yahoo! Answers based on the $STRICT$ method, $r_q$ of question $q$ in $QA\text{-}dataset$ is the *average* of the ranking positions of the *first relevant answer* to each question $Y_{Q_a}$, $Y_{Q_b}$, …, $Y_{Q_x}$ in $YA\text{-}Qs$ of $q$. This baseline simulates a user who "follows" the ranking of the retrieved questions and answers given by Yahoo! Answers and examines retrieved question threads and their corresponding answers in the order they were originally ranked.

**MRR-RR (Round Robin)**

In computing the MRR score of Yahoo! Answers as in Equation 9 using the $RR$ method, $r_q$ of question $q$ in $QA\text{-}dataset$ is computed using $YA\text{-}Qs$ of $q$ as follows: the $RR$ method treats the first answer of $Y_{Q_a}$ as the *first* answer to $YA\text{-}Qs$, the first answer of $Y_{Q_b}$ as the *second* answer to $YA\text{-}Qs$, and so on. Thereafter, $r_q$ is defined as the ranking position of the $first\ relevant\ answer$ among all the ranked answers in the ordered list. This baseline simulates a "jumpy" user who believes that answers that come $first$, no matter to which questions, are always $better$, and thus jumps between question threads examining the *top-ranked answer* for each question thread in the order of the original ranking.

The variants for $MAP$ and $P@K$ on ranked answers retrieved by Yahoo! Answers are computed in the same manner as the variants of $MRR$.

## 4.3 Performance Evaluation

In this section, we evaluate the *question-matching* and *answer-ranking* strategies of $QAR$ and compare the performance of these proposed strategies with the ones adopted by Yahoo! Answers.

### 4.3.1 Accuracy of Question-Matching

We determine the accuracy of $QAR$'s question-matching strategy based on the ratio of (up to top-10) archived questions $AQ$ (among the ones in $YA\text{-}Data$) extracted by $QAR$ in response to each question $Q$ in $QA\text{-}dataset$ that were identified as relevant by the independent appraisers who $matched$ the information needs specified in $AQ$ and $Q$. According to the experimental results, $QAR$ achieves an average of 70% accuracy in matching archived questions. Furthermore, $QAR$'s question-matching strategy, which is based on *word-correlation factors*, outperforms its counterpart adopted by Yahoo! Answers, which is based on *exact-keyword matching*, by 26% (see Figure 11). The average question-matching accuracy achieved by Yahoo! Answers, which is 44% and is based on the judgments of the same group of individual appraisers who evaluated $QAR$'s matched questions, indicates that on the average *4* out of 10 questions retrieved by Yahoo! Answers for a question $Q$ in $QA\text{-}dataset$ are the same or related to $Q$, as opposed to the *7* out of 10 retrieved by $QAR$.

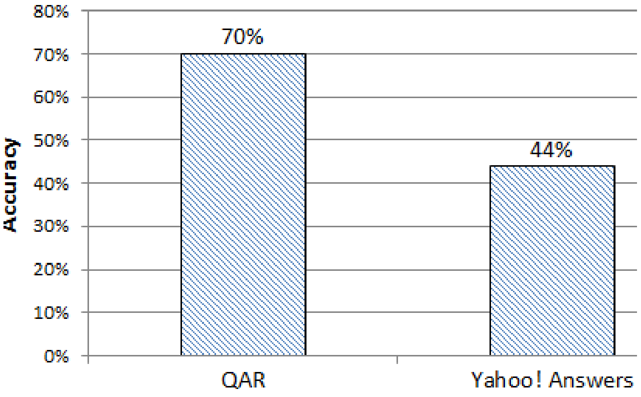


**Figure 11: Accuracy ratios of $QAR$ and Yahoo! Answers in identifying questions that are the same or semantically similar to each question in $QA\text{-}dataset$**

We have observed that out of the 300 $QA\text{-}dataset$ test questions used in the empirical study, $QAR$ found at least one (relevant) match for 30 more questions in $QA\text{-}dataset$ than Yahoo! Answers did. In addition, 29% of the questions (out of a total of 31) for which $QAR$ found no match while Yahoo! Answers did are questions (and their corresponding answers) that were posted (up till January 11, 2011) on Yahoo! Answers after its smaller subset $YA\text{-}Data$ was created in October 2007, which is used by $QAR$ for question answering, a disadvantage for $QAR$.

### 4.3.2 Accuracy of Answer-Ranking

We evaluate the *answer-ranking* strategy of $QAR$ and compare its performance with Yahoo! Answers' counterpart using the MRR, P@K ($K \in \{1, 5, 10\}$), and MAP scores of $QAR$ and Yahoo! Answers, respectively. Note that the metrics for Yahoo! Answers were computed using the three alternative strategies, i.e., $MAX$, $STRICT$, and $RR$, presented in Section 4.2.3. Each of the metric scores were computed using the ranked archived answers to the top-10 questions in $YA\text{-}Data$ (Yahoo! Answers, respectively) retrieved by $QAR$ (Yahoo! Answers, respectively) for each test question in $QA\text{-}dataset$[10].

The average MRR score of $QAR$, which is 0.58 (as shown in Figure 12), reflects that on an average a $QAR$ user is required to

[10] Recall that if an answer $A$ (retrieved by either $QAR$ or Yahoo! Answers) matches the *answer pattern* defined by TREC (is labeled as relevant by an independent appraiser, respectively) for a question $Q$ in $QA\text{-}dataset$ provided by TREC (TAC, respectively), then $A$ is treated as $relevant$ to $Q$.

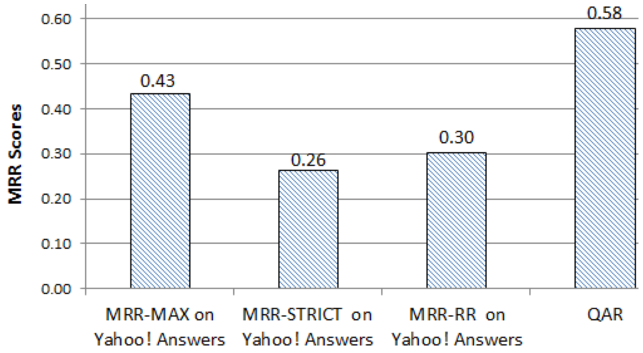


**Figure 12: (Average) MRR scores achieved by $QAR$ and Yahoo! Answers**

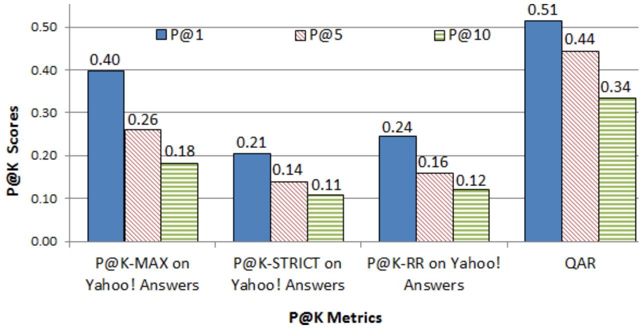


**Figure 13: (Average) P@K scores achieved by $QAR$ and Yahoo! Answers**

scan through less than *two* ($\cong \frac{1}{0.58}$ = 1.72) retrieved ranked answers before locating one that satisfies the information need expressed in his/her question. Users of Yahoo! Answers, on the other hand, are expected to access (on an average) *three* ($\cong \frac{1}{0.43}$ = 2.32), *four* ($\cong \frac{1}{0.26}$ = 3.85), and *four* ($\cong \frac{1}{0.30}$ = 3.33) answers before locating a relevant one according to the MRR-MAX, MRR-STRICT, and MRR-RR values, respectively of Yahoo! Answers.

Figure 13 shows the $P@K$ ($K \in \{1, 5, 10\}$) values, each of which estimates the number of relevant answers that appear in the top-$K$ results retrieved by $QAR$ or Yahoo! Answers in response to a (user's) question (in $QA$-$dataset$). While $QAR$ achieves 0.51, 0.44, and 0.34 for P@1, P@5, and P@10, respectively, which indicate that on an average $QAR$ can locate $K$ ($\in \{1, 5, 10\}$) relevant answers among the top-$K$ ranked answers to a new user's question $Q$ 43% (= $\frac{0.51+0.44+0.34}{3}$) of the time, Yahoo! Answers (when considering P@$K$-MAX which yields the highest P@$K$ score for Yahoo! Answers) accomplishes the same task for an average of 28% (= $\frac{0.40+0.26+0.18}{3}$) of the time. Based on the P@$K$ values in Figure 13, we claim that $QAR$ consistently outperforms Yahoo! Answers in terms of retrieving relevant answers to a question at the top-$K$ position.

We have also compared the MAP scores of $QAR$ and Yahoo! Answers using questions in $QA$-$dataset$. The average MAP score of $QAR$, which is 0.48 (as depicted in Figure 14), shows that on the average *five* (out of the top-10, if they exist) archived answers retrieved by $QAR$ to a new user's question $Q$ are *relevant*, whereas the *best* MAP score (i.e., MAP-MAX) of Yahoo! Answers, which is 0.19, shows that on an average Yahoo! Answers retrieves at least *three less* relevant answers to a user's question than $QAR$. Hence,

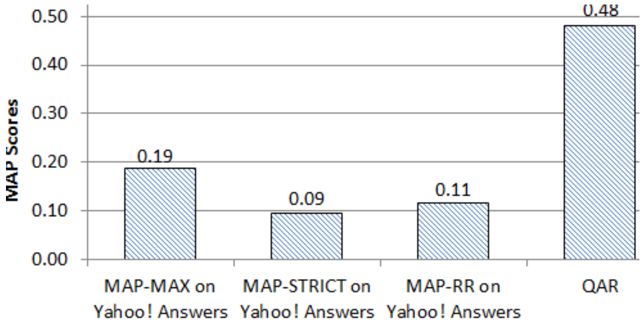


**Figure 14: (Average) MAP scores achieved by $QAR$ and Yahoo! Answers**

not only $QAR$ retrieves a $larger$ number of relevant answers in response to a user's question, but it also positions the relevant answers $higher$ in the ranking than Yahoo! Answers. As a result, $fewer$ answers are expected to be $examined$ or $accessed$ by a $QAR$ user than a Yahoo! Answers user in finding relevant ones.

## 5. CONCLUSIONS

We have introduced $QAR$, a Community Question Answering (CQA) refinement system that outperforms Yahoo! Answers in locating archived answers, if they exist, that satisfy the information need expressed in a new user's question $Q$. $QAR$ applies a *blocking* approach along with a simple, yet effective *question-matching* strategy based on word-correlation factors to identify the set of questions $QS$ that are the same, or related to, $Q$ among the millions provided by Yahoo! Answers. Thereafter, $QAR$ ranks each archived answer $A$ to its corresponding question $Q_A$ in $QS$ using an *answer-ranking* strategy, which is based on the similarity between $A$ and $Q$, as well as between $A$ and $Q_A$, and the length of $A$. The top-10 ranked answers, if there are any, are treated as answers to $Q$.

In developing $QAR$, we have solved many of the problems that currently affect CQA users, which include (i) receiving no answers at all to a new question $Q$ and (ii) waiting days for other CQA users to post answers to $Q$. Moreover, unlike existing CQA systems (such as Yahoo! Answers), $QAR$ does not impose an exact-matching constraint between (words in) CQA questions and $Q$, and thus retrieves (questions and) answers that are *relevant* to $Q$ even if they do not use exactly the same wordings as $Q$. Furthermore, $QAR$ retrieves ranked relevant answers to $Q$ without requiring its users to browse through CQA archived questions that are matched by CQA systems with respect to $Q$, which significantly minimizes the users' time and efforts involved in searching for answers to $Q$.

We have evaluated $QAR$ using (i) a set of 300 questions provided by the Text Retrieval Conference (TREC) and the Text Analysis Conference (TAC) as new user's questions, and (ii) more than four million questions and their corresponding answers extracted from Yahoo! Answers which serve as the source of $QAR$'s questions and answers. The conducted experiments have verified the $accuracy$ of $QAR$ in selecting $questions$ most similar to a user's question $Q$, in addition to its $effectiveness$ in retrieving relevant archived $answers$ to $Q$. Furthermore, we have compared the performance of $QAR$ with the one of Yahoo! Answers, and we have demonstrated that $QAR$'s strategy for locating archived answers is significantly more effective than the strategy adopted by Yahoo! Answers, a major community question-answering system.

# 6. REFERENCES

[1] Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding High-quality Content in Social Media. In: Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM), pp. 183–193. (2008)

[2] Bendersky, M., Croft, W.: Discovering Key Concepts in Verbose Queries. In: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pp. 491–498. (2008)

[3] Bian, J., Liu, Y., Agichtein, E., Zha, H.: Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media. In: Proceedings of the International Conference on World Wide Web (WWW), pp. 467–476. (2008)

[4] Bian, J., Liu, Y., Agichtein, E., Zha. H.: A Few Bad Votes Too Many?: Towards Robust Ranking in Social Media. In: Proceedings of the International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), pp. 53–60. (2008)

[5] Brants, T., Franz, A.: Web IT 5-gram Version 1 (www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13). (2006)

[6] Cao, X., Cong, G., Cui, B., Jensen, C., Zhang, C.: The Use of Categorization Information in Language Models for Question Retrieval. In: Proceedings of the ACM Conference on Information and Knowledge Management (CIKM), pp. 265–274. (2009)

[7] Christen, P.: Automatic Record Linkage Using Seeded Nearest Neighbor and Support Vector Machine Classification. In: Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 151–159. (2008)

[8] Church, K., Gale, W.: Poison Mixtures. Natural Language Engineering. 1(2), 163–190 (1995)

[9] Croft, W., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison Wesley, (2010)

[10] Gustafson, N., Ng, Y.-K.: Augmenting Data Retrieval with Information Retrieval Techniques by Using Word Similarity. In: Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB), pp. 163–174. (2008)

[11] Hoscher, C., Strube, G.: Web Search Behavior of Internet Experts and Newbies. Computer Networks: The International Journal of Computer and Telecommunications Networking. 33, 337–346 (2000)

[12] Jeon, J., Croft, W., Lee, J.: Finding Similar Questions in Large Question and Answer Archives. In: Proceedings of the ACM Conference on Information and Knowledge Management (CIKM), pp. 84–90. (2005)

[13] Jeon, J., Croft, W., Lee, J., Park. S.: A Framework to Predict the Quality of Answers with Non-textual Features. In: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pp. 228–235. (2006)

[14] Judea, P.: Probabilistic Reasoning in the Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann. (1988)

[15] Kelley, R.: Blocking Considerations for Record Linkage Under Conditions of Uncertainty. In: Proceedings of Social Statistics Section, pp. 602–605. (1984)

[16] Koberstein, J., Ng, Y.-K.: Using Word Clusters to Detect Similar Web Documents. In: Proceedings of the International Conference on Knowledge Science, Engineering and Management (KSEM), pages 215–228. 2006.

[17] Lee, C., Rodrigues, E., Kazai, G., Milic-Frayling, N., Ignjatovic. A.: Model for Voter Scoring and Best Answer Selection in Community Q&A Services. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT), pp. 116–123. (2009)

[18] Liu, Y., Agichtein, E.: On the Evolution of the Yahoo! Answers QA Community. In: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pp. 737–738. (2008)

[19] Luger, G.: Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 6th Ed. Addison Wesley. (2009)

[20] Manning, C., Raghavan, P., Schutze, H.: Introduction to Information Retrieval, Cambridge University Press. 2008.

[21] Pera, M.S., Lund, W., Ng, Y.-K.: A Sophisticated Library Search Strategy Using Folksonomies and Similarity Matches. Journal of the American Society for Information Science and Technology (JASIST). 60(7), 1392–1406 (2009)

[22] Sahami, M., Heilman, T.: A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets. In: Proceedings of the International Conference on World Wide Web (WWW), pp. 377–386. (2006)

[23] Salton G., Buckley C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management: an International Journal. 24(5), 513–523 (1988)

[24] Spink, A., Ozmutlu, S., Ozmutlu, H., Jansen, B.: U.S. versus European Web Searching Trends. ACM SIGIR Forum. 36(2), 32–38 (2002)

[25] Suryanto, M., Lim, E., Sun, A., Chiang, R.: Quality-aware Collaborative Question Answering: Methods and Evaluation. In: Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM), pp. 142–151. (2009)

[26] Tu, X., Wang, X., Feng, D., Zhang, L.: Ranking Community Answers via Analogical Reasoning. In: Proceedings of the International Conference on World Wide Web (WWW), pp. 1227–1228. (2009)

[27] Wang, K., Ming, Z., Chua, T.: A Syntactic Tree Matching Approach to Finding Similar Questions in Community-based QA Services. In: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pp. 187–194. (2009)

[28] Wang, X., Tu, X., Feng, D., Zhang, L.: Ranking Community Answers by Modeling Question-Answer Relationships via Analogical Reasoning. In: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pp. 179–186. (2009)

[29] Xue, X., Jeon, J., Croft, W.: Retrieval Models for Question and Answer Archives. In: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pp. 475–482. (2008)

[30] Yahoo! Webscope Dataset.: L6-Yahoo! Answers Comprehensive Questions and Answers version 1.0. http://research.yahoo.com/Academic_Relations. 2009.

[31] Zhou, Y., Croft, W.: Query Performance Prediction in Web Search Environments. In: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pp. 543–550. (2007)