Clustering Similar Questions In Social Question Answering Services

Mohan John Blooma
Center of Commerce, Royal Melbourne Institute of Technology, Ho Chi Minh City, Vietnam, blooma.john@rmit.edu.vn

Jayan Chirayath Kurian
Center of Commerce, Royal Melbourne Institute of Technology, Ho Chi Minh City, Vietnam, jayan.kurian@rmit.edu.vn

Follow this and additional works at: http://aisel.aisnet.org/pacis2012

Recommended Citation
http://aisel.aisnet.org/pacis2012/160

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2012 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
CLUSTERING SIMILAR QUESTIONS IN SOCIAL QUESTION ANSWERING SERVICES

Mohan John Blooma, Center of Commerce, Royal Melbourne Institute of Technology, Ho Chi Minh City, Vietnam
Jayan Chirayath Kurian, Center of Commerce, Royal Melbourne Institute of Technology, Ho Chi Minh City, Vietnam

Abstract
Social Question Answering (SQA) services are defined as dedicated platforms for users to respond to other users’ questions, resulting in the building of a community where users share and interactively give ratings to questions and answers (Liu et al., 2008). SQA services are emerging as a valuable information resource that is rich not only in the expertise of the user community but also their interactions and insights in the form of user comments and ratings. In SQA services each user interaction is different and since there are a variety of complex questions, identifying similar questions for re-using answers is difficult. Scholarly inquiries have yet to dovetail into a composite research stream in identifying similar questions by harnessing the information richness in SQA services. This paper aims to develop a quadripartite graph-based clustering (QGC) approach by harnessing relationship of a question with common answers and associated users. It was found that QGC approach outperformed other baseline clustering techniques in identifying similar questions in SQA corpora. We believe that these findings can serve to guide future developments in the reuse of similar question in SQA services.

Keywords:
Social question answering, Text mining, Graph-based clustering
1 INTRODUCTION

A SQA service is defined as a tool for users ask, answer and rate other users’ questions and answers (Gazan, 2011). In recent years, SQA services like Yahoo! Answers, Naver, and AnswerBag have become very popular, attracting a large number of users who seek and contribute answers to a variety of questions on diverse subjects (Wang et al., 2009). Social computing applications such as wikis and blogs provide comments and opinions but may not solicit responses. However, responses from users contributing answers to questions form the backbone of a successful SQA service. These services are dedicated platforms for user-oriented Question Answering (QA) and result in building up a community where users share and interactively give ratings and comments to questions and answers. Hence, SQA services are emerging as a valuable information resource that is rich not only in the expertise of a user community but also in the community’s interactions and insights. Therefore, this information resource could be harnessed to build an automatic QA system.

However, a major challenge in automating SQA services lies in identifying similar questions due to the nature of user-generated questions. There is a need to find questions in a corpus that were semantically similar to a user’s newly posed question (Jeon et al., 2005a,b). This facilitates the retrieval of high-quality answers that are associated with similar questions identified in the corpus, reducing the time lag associated with a SQA system.

Traditional QA systems handle single-sentence questions that are particularly fact-based. Single-sentence questions are defined as questions composed of one sentence. Since 2006, TREC and other QA research groups have started to focus on complex and interactive questions (Dang et al., 2006). However, they are still in the early stages of research. SQA services are rich in multiple-sentence questions, which are defined as questions composed of two or more sentences: For example, “I would like to download some good quality clips. Please recommend a good site to get good clips and also how to go about it?” This could be formed as a simple direct question given as, “How to download videos from YouTube?”.

Due to the complex nature of questions posed in SQA services, it is not an easy task to identify similar questions by measuring their semantic similarities. This is because two questions having the same meaning may use entirely different wording. Similarity measures developed for document retrieval work poorly when there is little word overlaps (Tamura et al., 2005). Hence, there is a need to work on more comprehensive methods that make use of available metadata to identify similar questions (Bian et al., 2009). Moreover, as there is an increasing amount of data collected in the SQA corpora, identifying semantically similar questions from the corpora would be more challenging and rewarding for the on-going research in this domain.

Various techniques, like classification and query expansion, have been used for understanding simple, fact-based questions, and for grouping similar questions. However, very little research has been carried out on finding similar questions from SQA corpora (Tamura et al., 2005). Hence this research gap could be filled by proposing clustering methods to identify similar questions in SQA corpora. In previous studies, complex questions from SQA corpora were classified by simplifying the question and then applying question classification algorithms (Tamura et al., 2005). Another approach used in this context is a statistical translation based model to find semantically similar questions based on their answers (Jeon et al., 2005b). Identifying similar questions by clustering them into similar categories could be used for improving answer retrieval. It is widely accepted as a measure to increase efficiency and accuracy of retrieval systems (Tombros et al., 2002). However, few researchers have used clustering as a means to identify similar questions, although query clustering is a well-researched domain (Beeferman & Berger, 2000; Leung et al., 2008).

Drawing insights from previous studies and based on the identified research gap, this paper aims to propose and evaluate a technique for clustering questions in SQA corpora. Clustering questions will enable similar questions to be logically grouped. Clustering methods that use word similarities
between questions would be insufficient in this context for several reasons. First, as argued by (Leung et al., 2008), clustering methods using word similarities between questions would be insufficient due to the lexical mismatch problem between questions posed in SQA services. Second, due to the open nature of SQA services, questions are posed in a free form to the user community (Beeferman & Berger, 2000; Tamura et al., 2005). Therefore, this paper focuses on using the relationship of a question with common answers and users, as additional criteria for identifying similar questions.

The remainder of this paper is organized as follows. The next section gives an overview of previous literature on clustering. Thereafter, the development of the proposed clustering algorithm is presented. This is followed by the explanation of the experiments used to evaluate the performance of the proposed clustering algorithm. We then discuss the research findings and the opportunities for future work. This paper is concluded with a summarization of the findings and its implications.

2 LITERATURE REVIEW

2.1 Text Clustering

Clustering is defined as the grouping together of similar objects (Dhillon, 2001). Works related to clustering similar questions are presented in two strands: document clustering and query clustering. Each of these strands and the problems encountered in adapting these clustering methods to questions in SQA corpora are detailed below.

2.1.1 Document Clustering

A starting point for applying clustering algorithms to document collection is to adopt the vector space model (Salton & McGill, 1983). Existing document clustering methods include agglomerative clustering (Walter et al., 2008), the partitional k-means algorithm (Angelova & Sierdorfer, 2006) and projection-based methods including least squares approximation (LSA) (Kim & Park, 2008). The underlying assumption is that words that typically appear together should be associated with similar concepts. In document clustering a document can be represented by a relatively large number of content words. However, questions submitted to a SQA service are usually of varying length and a clustering approach using solely keywords in questions will not be effective (Jeon et al., 2005b). Moreover, two questions that have the same meaning may use very different wordings. Therefore, similarity measures developed for document retrieval worked poorly when there was little word overlap. This is the same for traditional sentence distance measures such as the Jaccard coefficient and the overlap coefficient (Salton & McGill, 1983).

2.1.2 Query Clustering

Query clustering is a branch of text clustering that focuses on the length and lexical disagreement problems (Wen et al., 2002; Leung et al., 2008). Queries are typically very short and in many cases it is hard to deduce the semantics from the queries themselves. Therefore, keywords alone do not provide a reliable base for clustering queries effectively. To resolve the disadvantages of keyword based query clustering, newer research focused on additional criteria. One criterion is hyperlinks between documents, based on the hypothesis that hyperlinks connect similar documents. Beeferman and Berger (2000) used an agglomerative clustering algorithm (i.e., BB’s algorithm) to exploit query-document relationships using click-through data. In agglomerative clustering, objects were initially assigned to their own cluster and then pairs of clusters were repeatedly merged until the whole dendogram was formed. One disadvantage of their approach was that the algorithm was content-independent, in the sense that it exploited only query-document links to discover similar queries and similar documents, without examining keywords in queries or documents. Another major problem was that the number of common clicks on URLs for different queries was limited. This is because different queries will likely retrieve different result sets in different ranking orders. Due to these disadvantages of using hyperlinks, studies emphasized cross references between documents and queries in query-
document clustering (Wen et al., 2002). The idea behind this type of clustering is if queries’ often lead to similar documents, then those queries are similar, to some extent. However, they do not consider the content. To alleviate this problem, a more recent study (Leung et al., 2008), introduced the notion of concept-based graphs, by considering concepts extracted from web-snippets, and adapted method proposed by Beeferman and Berger, (2000) to this new context. However, their work also neglected word similarity between queries. Nevertheless, one disadvantage of the above studies on query clustering is that the algorithms used were content-independent, in the sense that bipartite graph-based clustering exploited only query-document/concept links to discover similar queries without examining the keywords in the queries or the documents. To overcome this shortcoming, the proposed quadripartite graph-based clustering extends the relationship-based similarity by combining it with content-based similarity as detailed in the next section.

2.2 Graph-based Clustering

Mapping web-based social interactions onto a graph represents a classical example of the applied theory of complex networks (Boccaletti et al., 2006). Two typical features of these networks with different origins are community structure (Arenas et al., 2004) and assortative mixing (Newman, 2002), both of which have a clear analogy in real-life experiences. As a multipartite network representation of web-based social interactions, both interacting agents (humans, users of the Web portals) and subjects of their interactions (music, movies, books, postings) in a social network are represented by nodes on the graph and their mutual connections can be analysed in detail. Studies by Lambiotte and Ausloos, (2005a; 2005b) are examples of social connections related to music, where communities related to music genres have been detected. Specifically, studies on graph-based clustering could be reviewed depending on various applications related to bipartite and tripartite networks.

2.2.1 Bipartite Networks

Bipartite networks are composed of two kinds of nodes. Some different types of nodes used in bipartite networks are query-document (Rege et al., 2008), query-URL (Li et al., 2007), user-movie (Grujic et al., 2009) and question-answer (Bian et al., 2009). Bipartite networks have been considered for various applications, including text clustering (Li et al., 2008), ontology mapping (Chen & Fonsecca, 2003), identifying user communities (Grujic et al., 2009), extracting verb synonyms (Takeuchi, 2008) and recognizing reliable users and content in social media (Bian et al., 2009).

Studies related to text clustering used bipartite graphs for clustering queries using hyperlinks (Beeferman & Berger, 2000), co-clustering documents and words (Rege et al., 2006), query clustering (Li et al., 2007) and concept-based query clustering (Leung et al., 2008). The underlying assumption for these studies was that concepts that typically appear together should be similar concepts. Based on this assumption, Beeferman and Berger, (2000) considered hyperlinks to cluster queries, while Leung et al. (2008) considered commonly occurring concepts in web-snippets. However, neither of these studies considered query-content similarity or user similarity. A recent study on identifying high-quality content and users in SQA services used a mutually coupled bipartite network (Bian et al., 2009). Interactions in SQA were considered as composite bipartite graphs and the mutual reinforcement between the connected entities was exploited in each bipartite graph to compute their respective quality and reputation scores.

2.2.2 Tripartite Networks

Tripartite networks are composed of three kinds of nodes. Some different types of nodes used in tripartite networks are user-resource-tags (Lu et al., 2009) and visual feature - Web image - related text (Rege et al., 2006). Tripartite networks have been considered for various applications such as collaborative tagging (Lambiotte & Ausloos, 2006), Web clustering (Lu et al., 2009), and Web image
clustering (Rege et al., 2008). Networks formed using tripartite graphs are superior to bipartite networks as they consider the possibility of correlations among three kinds of nodes.

Among studies related to clustering, Lambiotte and Ausloos, (2006) used users, resources and tags as nodes in a tripartite network for collaborative tagging. Similar nodes were also used by (Lu et al., 2009) to investigate how to enhance Web clustering by leveraging the tripartite network of social tagging systems. Rege et al. (2008) proposed a tripartite network of visual and textual features of images for efficient Web image clustering. An example of a visual feature is the colour histogram. They were able to address the semantic gap between visual features and high level semantic concepts, which was considered to be one of the shortcomings of Web image clustering.

Thus, the studies reviewed above form the foundation for the proposed quadripartite graph-based clustering techniques built in this paper for clustering similar questions. The proposed quadripartite clustering is detailed in the following section.

3 QUADRIPARTITE GRAPH-BASED CLUSTERING

3.1 Question-Answer-Asker-Answerer Quadripartite Network

Using shared knowledge acquired from SQA services, a quadripartite network was constructed with concepts extracted from the best answer selected by the asker, asker profile and answerer profile related to a question. This network is termed a Question-Answer-Asker-Answerer quadripartite network. The quadripartite structure of a Question-Answer-Asker-Answerer network differs fundamentally from the bipartite structure of hyperlink or concept-query graph (Leung et al., 2008) in that it exploits not only the relationships between questions and answers but also between askers and answerers who contributed to these questions and answers. The basic assumptions are as follows: questions can be clustered on the basis of similar answer concepts. Askers posing similar questions tend to have similar information needs. Likewise, answerers offering similar answer concepts share similar expertise. Thus, there exists an intricate network of relationships among questions, answer concepts, askers and answerers as discussed in the earlier work by Blooka et al., 2011.

As shown in Figure 1, the relationships among these four entities, namely, questions (q), answer concepts (ac), askers (as) and answerers (an) collectively form a quadripartite network. Each time an asker asks a question and selects a best answer, a quadripartite relationship is built among the question, its answer concepts extracted from the question’s respective best answer, its asker and its answerer. Answer concepts are obtained from answers by extracting the important phrases.

![Figure 1. An example of quadripartite network of SQA](image-url)
3.2 Quadripartite Graph-based Clustering (QGC)

Similar to the algorithm used by Beeferman and Berger, (2000) and Leung et al. (2008) the proposed QGC technique is composed of two steps, quadripartite graph construction using extracted concepts, askers and answerers followed by agglomerative clustering on the quadripartite graph constructed.

Algorithm 1 Quadripartite Graph (QG) Construction

Input: Question (Q) and its related answer concepts (AC), asker (AS) and answerer (AN) relationships are collectively termed as $Q_{relation}$.

Output: A Question-Answer-Asker-Answerer Quadripartite Graph $QG$

1) Obtain the set of unique questions $Q = \{ q_1, q_2, q_3, \ldots \}$ from $Q_{relation}$.
2) Obtain the set of unique answer concepts $AC = \{ ac_1, ac_2, ac_3, \ldots \}$ from answers $A = \{ a_1, a_2, a_3, \ldots \}$ in $Q_{relation}$.
3) Obtain the set of unique askers $AS = \{ as_1, as_2, as_3, \ldots \}$ from $Q_{relation}$.
4) Obtain the set of unique answerers $AN = \{ an_1, an_2, an_3, \ldots \}$ from $Q_{relation}$.
5) Total number of Nodes in $QG = Q \cup AC \cup AS \cup AN$, where $Q$, $AC$, $AS$ and $AN$ are four sides in $QG$.
6) If an answer $a_i$ answered by the answerer $an_l$, is marked as best answer by the asker $as_k$ for the question $q_l \in Q$, create the following edges:
   i. an edge $e_{q,a} = (q_l, ac_j)$ in $QG$, for all answer concepts $ac_j$ in answer $a_i$ to question $q_l$.
   ii. an edge $e_{a,a} = (a_i, as_k)$ in $QG$, where $as_k$ is asker of question $q_l$.
   iii. an edge $e_{q,a} = (q_l, an_l)$ in $QG$, where $an_l$ is answerer of answerer $a_i$.
   iv. an edge $e_{a,a} = (ac_j, as_k)$ in $QG$, for all answer concepts $ac_j$ in answer $a_i$ to question $q_l$ asked by asker $as_k$.
   v. an edge $e_{a,a} = (ac_j, an_l)$ in $QG$, for all answer concepts $ac_j$ in answer $a_i$ answered by answerer $an_l$.
   vi. an edge $e_{a,a} = (as_k, an_l)$ in $QG$.

Using extracted answer concepts from the best answer marked for a question, asker of the question and answerer of the best answer, a quadripartite graph (QG) is constructed. In the quadripartite graph, the first side of the vertices corresponds to unique questions, the second side to unique answer concepts, the third side to a unique set of askers and the fourth side to a unique set of answerers whose answers where selected as best answers. If a question was asked by the asker and received a best answer from an answerer, six different types of links could be formed in the quadripartite graph.

After construction of the quadripartite graph, agglomerative clustering algorithm is applied to obtain clusters of similar questions, answer concepts, askers and answerers. The noise-tolerant similarity function proposed in (Leung et al., 2008) was adapted for finding similar vertices on the quadripartite graph $QG$.

Thus, based on the above quadripartite graph and the similarity functions, QGC using relationship similarity measure was performed. Agglomerative clustering is used to merge the most similar pair of vertices using their respective similarly measures. This would iteratively merge the most similar pair of question vertices followed by answer concept vertices, asker vertices and answerer vertices until the termination condition is reached. The termination condition is given as a particular threshold for the similarity value. $\alpha$, $\beta$, and $\gamma$ are termed as relationship parameters.

QGC was further enhanced using a combined similarity measure by including question word similarity. This enhancement is required because similarity measures based on question content words and the Question-Answer-Asker-Answerer relationship represent two different points of view. First, content-based similarity measures tend to cluster questions with the same or similar terms, but similar terms could be used to represent different requirements because of the ambiguity of words. Second, similarity measures based on the Question-Answer-Asker-Answerer relationship tend to cluster questions related to the same or similar topics. However, an asker or answerer might have more than one topic of interest. Answer concepts might be used in different contexts. So, questions with different meanings could lead to answers containing the same answer concepts or the same askers and answerers. Since a user’s information needs may be partially captured by each of the above criterion, the proposed quadripartite algorithm based on the Question-Answer-Asker-Answerer relationship is
extended to consider the question words based on the assumption that if two questions contain the same or similar terms, they have the same or similar meaning. Moreover, although Leung et al. (2008) used content-independent bipartite graph-based clustering, it was one of their shortcomings. Hence clustering using a combined similarity was also performed.

4  METHODOLOGY

A detailed explanation of the methodology used to test the performance of the proposed clustering is given in this section. The process involved in data collection is first explained which is followed by the explanation of the performance evaluations used. The experiments conducted for identifying the best question clustering technique is then detailed.

4.1  Data Extraction

This study was based on the data collected from Yahoo! Answers. Yahoo! Answers represents an online community where users could ask or answer questions on a wide range of topics like politics, sports, health etc. Yahoo! Answers was used as the dataset for this study for its popularity, richness in content and metadata. In particular, to identify similar questions, data collected from Yahoo! Answers are deemed suitable for two reasons. First, because of the range in subjective orientation of questions in Yahoo! Answers, there is an urge for new techniques to identify similar questions (Li et al., 2007). The questions posted in Yahoo! Answers are typically complex, subjective, and rely on human interpretation to understand the corresponding information needs. Second, the questions are usually of varying length, ill-phrased or vague. Hence, analysis of ambiguous questions is a particularly difficult task.

Data was collected from Yahoo! Answers by scripting a crawler in C# developed for this research. The crawler was used to harvest questions, answers and related user profiles from four different categories. The four categories were selected from among the most popular categories in Yahoo! Answers. They were Arts & Humanities, Business & Finance, Computers & Internet, and Science & Mathematics. Questions harvested were restricted to be taken from the resolved section in Yahoo! Answers. The resolved section was selected because a question in this section will contain a best answer preferred by the asker along with all other answers obtained for the question. Data collection was performed during first week of May 2009. Crawling was performed to collect most recent questions posed in the resolved section. The selection was restricted to the questions posed within a time span of 7 days.

In Yahoo! Answers, the main question gives the title of the question and was used for the clustering experiments because it contained the essence of the question. Inclusion of sub-questions would require complex natural language processing techniques to interpret and summarize its essence. Moreover previous studies assumed that main questions provide enough contextual information for understanding users’ information needs (Jeon et al., 2005a). Best answers were used for extracting answer concepts. The dataset was manually checked for valid data in all fields and irrelevant data were removed. For example, questions like “Please help” and answers such as “no” were expunged. The questions and answers were also processed to remove all unwanted special characters used, such as emoticons. The dataset used for clustering questions, after pre-processing using the above measures, contained a total of 5,733 questions from four popular categories.

4.2  Evaluating Performance of Question Clusters

After obtaining clusters of questions, the performances of the different clustering algorithms were assessed. Since quadripartite graph-based clustering is an unsupervised process, there are no predefined classes against which the validity of the clusters can be judged. This makes it necessary to employ metrics that will give a suitable indication of the quality of clusters produced by the algorithm. Precision, Recall and F-Measure are the more commonly used performance metrics in this case.
(Carpineto et al., 2009). Precision is measured as the ratio of the number of similar questions to the total number of questions in a cluster. Recall is measured as the ratio of the number of similar questions in the current cluster to the total number of all similar questions for a question set. F-Measure is the harmonic mean of recall and precision. Clusters are compared with reference to external knowledge against some predefined set of desirable qualities. In this case, the quality of clusters was compared against judgments made by two human evaluators. Two evaluators were required because the judgment of the relatedness of two questions is subjective and there is room for personal biases. These biases can be eliminated to a great extent by taking an aggregate of the two evaluators’ appraisals.

Human evaluators evaluated 100 random samples of clusters from each of the thresholds. Each cluster given for evaluation had two or more questions. The evaluators were asked to identify the questions in a cluster that had same meaning and mark them against a check box. Since, in some cases it is difficult to correctly understand the intention of the user, evaluators just made the best guess. Each of the evaluators reviewed the questions independently. Finally, the metrics were calculated for each evaluator’s sample and the two figures obtained were averaged to get the final performance metrics. Also, the degree of agreement between evaluators was calculated. Cohen’s Kappa statistic was adopted to measure the degree of agreement between two evaluators’ judgments (Cohen, 1960). A Cohen Kappa value of 0.82 was obtained as the measure of agreement between the two evaluators for their manual evaluation of clusters.

4.3 Experiments on Quadripartite Graph-based Clustering

4.3.1 Effect of using a relationship similarity measure

For quadripartite graph-based clustering using a relationship similarity measure \(QR\_Sim\), there is a need to identify the best combination of weights for relationship parameters \(\alpha\), \(\beta\), and \(\gamma\). Hence, \(QR\_Sim_1\), \(QR\_Sim_2\), and \(QR\_Sim_3\) takes the values \{0.80, 0.10, 0.10\}, \{0.34, 0.33, 0.33\} and \{0.20, 0.40, 0.40\} respectively. Thus, three experiments were conducted to determine which combination of relationship parameters generated the best performance. Here, the performance metrics described in previous section were adopted to compare clusters.

Table 1 compares the quadripartite graph-based clustering technique using a relationship similarity measure taking the impact of different weights for relationship parameters \(\alpha\), \(\beta\), and \(\gamma\) into consideration. \(QR\_Sim_1\) \((\alpha = 0.80, \beta = 0.10, \gamma = 0.10)\) was found to perform the best of the three different combinations of weights used. \(QR\_Sim_1\) was found to perform best because it gave the highest values at all thresholds. \(QR\_Sim_1\) was found to perform best at a threshold value of 0.50. \(QR\_Sim_2\) gave equal weights for answer concepts, askers and answerers and hence did not form any clusters for the threshold value of 0.90, as it clustered questions that had exactly similar answers, askers and answerers. \(QR\_Sim_3\) also gave poor results because, although it gave similar questions, the same set of questions was clustered for all three threshold values. This was because very few questions had exactly the same askers and answerers and \(QR\_Sim_3\) gave importance to askers and answerers more than answer concepts. Hence, \(QR\_Sim_1\) was selected for the final experiment, comparing four different clustering techniques.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>(QR_Sim_1)</th>
<th>(QR_Sim_2)</th>
<th>(QR_Sim_3)</th>
<th>(QR_Sim_4)</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>64.3</td>
<td>100</td>
<td>100</td>
<td>70.2</td>
<td>33.3</td>
<td>33.3</td>
<td>67.1</td>
</tr>
<tr>
<td>0.70</td>
<td>65.2</td>
<td>100</td>
<td>100</td>
<td>59.3</td>
<td>20</td>
<td>20</td>
<td>62.1</td>
</tr>
<tr>
<td>0.90</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>47.2</td>
<td>0</td>
<td>20</td>
<td>64.1</td>
</tr>
</tbody>
</table>

Table 1 Quadripartite graph-based clustering result using a relationship similarity
4.3.2 Effect of using a combined similarity measure

For quadripartite graph-based clustering using a combined similarity measure, there is a need to identify the best combination of weights for the content parameter δ. Hence, $QRC_{Sim_1}$, $QRC_{Sim_2}$, and $QR_{Sim}$ take the value of δ as 0.20, 0.50 and 0.80 respectively. Thus three experiments were conducted to determine which value of δ generated the best performance. Each experiment was conducted for three thresholds as shown in Table 2. Table 2 compares different quadripartite graph-based clustering techniques using a combined similarity measure taking the impact of different values for content parameters into consideration. Taking all the performance metrics into consideration, $QRC_{Sim_1}$ (δ = 0.20) was found to perform the best of the three and thus was selected for the next experiment. $QRC_{Sim}$ was found to perform the best because it gave highest values for all the three thresholds.

<table>
<thead>
<tr>
<th>Weight</th>
<th>$QRC_{Sim_1}$</th>
<th>$QRC_{Sim_2}$</th>
<th>$QR_{Sim_1}$</th>
<th>$QR_{Sim_2}$</th>
<th>$QR_{Sim}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>62.3</td>
<td>100</td>
<td>100</td>
<td>80.5</td>
<td>33.3</td>
</tr>
<tr>
<td>0.70</td>
<td>70</td>
<td>0</td>
<td>100</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>0.90</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>56.7</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2 Quadripartite graph-based clustering results using a combined similarity

4.4 Comparison of Agglomerative, Bipartite, and Quadripartite Clustering

In order to evaluate the performance of quadripartite clustering, the results obtained were compared to two baseline methods. The first baseline method, agglomerative clustering ($AC_{Simbow}$ and $AC_{Simphrases}$), used question content similarity measures. The second baseline method used was bipartite graph-based clustering ($BR_{Sim}$) using a question-answer concept relationship similarity. The two baseline methods were compared with the proposed QGC using a relationship similarity measure ($QR_{Sim_1}$) and QGC using a combined similarity measure ($QRC_{Sim_2}$). Figure 3 gives the comparison of the best outputs for all the clustering techniques. The results were further analyzed and discussed below.

![Figure 3. Precision versus recall when comparing different clustering algorithms](image_url)

Legend:
- $AC_{Simbow}$: Agglomerative Clustering using the content similarity of a bag of words
- $AC_{Simphrases}$: Agglomerative Clustering using the content similarity of phrases
- $BR_{Sim}$: Bipartite graph-based clustering using similarity based on the question and answer concept relationship
- $QR_{Sim_1}$: QGC using a relationship similarity based with $\alpha = 0.80$, $\beta = 0.10$, and $\gamma = 0.10$
- $QRC_{Sim_2}$: QGC using a combined similarity based with $\delta = 0.20$

Results obtained shed new light in understanding the influence of the QGC technique in identifying similar questions from a SQA corpus. It was established in previous studies that questions leading to
similar concepts could be clustered as similar (Beeferman and Berger, 2000; Leung et al., 2008). However, the findings from the experiments discussed extend existing research by revealing the importance of combining question content similarity along with Question-Answer-Asker-Answerer relationship similarity in identifying similar questions. It is evident from Figure 3 that quadripartite graph-based clustering using a relationship similarity measure and a combined similarity measure both improved the baseline clustering techniques. On further analysis, three implications could be drawn from the findings as listed below.

First, quadripartite graph-based clustering using a relationship similarity measure was able to overcome the lexical mismatch in questions posed in SQA corpora. This was achieved by incorporating a Question-Answer-Asker-Answerer relationship similarity measure, and was found to perform better than agglomerative clustering using content similarity. Agglomerative clustering relied on common occurrences of content words, while quadripartite graph-based clustering used a relationship similarity measure relying on common occurrences of answer concepts, askers and answerers to identify similar questions. For example, “Health insurance and cosmetic surgery?” and “Health Insurance for non-citizens?” were found to be similar using agglomerative clustering. They had high word overlap but were of different concepts. However, quadripartite graph-based clustering using a relationship similarity was not able to cluster them because their answers, askers and answerers were different. Another example, “What is a deductible?” and “Need health insurance?” were identified as similar questions by quadripartite graph-based clustering using a relationship similarity although they were lexically different because they had similar answers as well as answerers. Hence quadripartite graph-based clustering using a relationship similarity was able to overcome the lexical gap in identifying similar questions posed in SQA services. (Lu et al., 2009) had similar results by using social tagging data to improve Web clustering. Leung et al. (2008) also incorporated a query-concept relationship to improve query clustering. However, when compared to the above two studies, quadripartite graph-based clustering using a relationship similarity measure clustered questions while previous studies clustered tagged resources and queries. This difference is important because questions posed are in natural language and lexical mismatch was one of the problems in identifying similar questions that the proposed clustering method was able to overcome.

Second, quadripartite graph-based clustering using a relationship similarity was able to incorporate user-attribution by relating questions and answers with their askers and answerers. This was achieved by incorporating a Question-Answer-Asker-Answerer relationship similarity measure and allowed the technique to perform better than bipartite graph-based clustering using question-answer concept relationship similarity. On further analysis of the results it was found that bipartite graph-based clustering resulted in questions that had similar answer concepts but questions were lexically as well as conceptually different and hence had reduced precision. This was one of the shortcomings of Beeferman and Berger, (2000) algorithm identified by Leung et al., (2008). To incorporate user-attribution, quadripartite graph represented similar questions as different nodes identified by its answer concepts, askers and answerers and effectively merged similar questions based on the relationship similarity measure. Moreover, in SQA services, few answerers gave general answers for questions repeatedly which led to clustering unrelated questions on the basis of similar answers. Hence it was evident that the use of answers alone was misleading to identify similar questions. For example, “Does it seem obscured to anyone but me, that hearing loss is not covered by medical insurance? Why?” and the question “Please help me with A & B supplement insurance?” got the same answer as “make use of the SE like Google or Yahoo to get some ideas first if you want to get the massive information, however if you do not want to spend so much time, here http://www.insuranceidea.info/free is a direct and good resource for your questions.” These two questions were clustered as similar by bipartite graph-based clustering as they had similar answers. However, quadripartite graph-based clustering using a relationship similarity measure was able to overcome this shortcoming by incorporating askers and answerers in addition to answers to incorporate user-attribution.

Third, by incorporating the combined similarity measure considering content as well as relationship, quadripartite graph-based clustering improved performance over three other clustering techniques,
quadrupartite graph-based clustering using a relationship similarity, bipartite graph-based clustering based on a question-answer concept relationship, and agglomerative clustering based on content. Wen et al., (2008) and Fu et al., (2004) are two studies that used a combined measure to check the quality of clustering queries and they obtained similar results. However, previous studies on a combined measure were still restricted to only query contents and user clicks. A recent study by Lu et al., (2009) used content words in resources and tags along with users to form a tripartite network for clustering. They demonstrated improved results when they combined the content and relationship similarity for clustering.

The proposed quadrupartite graph-based clustering used shared knowledge acquired from SQA services to identify similar questions based on the Question-Answer-Asker-Answerer network relationship. As illustrated in Figure 4, questions collected in a SQA corpus are clustered using the proposed quadrupartite graph-based clustering approach to identify similar questions. From the similar questions identified, a newly posed question could be paired with the best matching question and reuse its answers. If similar questions are available, this process eliminates the need for users to wait for other users to answer questions, and also the need to manually search or browse for similar questions from the resolved questions section available on SQA services.

![Figure 4. Automated SQA system using quadrupartite graph-based clustering](image)

5 CONCLUSION

This paper proposed and evaluated a novel clustering approach to identify similar questions in SQA corpora. The findings presented in this paper addressed the research objective by harnessing the shared knowledge garnered in SQA corpora to identify similar questions using the proposed quadrupartite graph-based clustering.

In general, there are two limitations for the use of quadrupartite graph-based clustering for identifying similar questions. The first limitation is the computational time complexity in quadrupartite graph-based clustering. Hence, as a part of future work, there is a need to optimize the computational time complexity. Avoiding recalculation of only merged vertices could help reduce the time (Beeferman and Berger, 2000). Another measure that could be used to reduce computational time complexity is to consider edges with greater than a predefined threshold in the graph (Chen & Fonsec, 2003). Second, this approach may become too slow for large and dynamic collections such as SQA corpora. Even though clustering is performed offline, there is a need to consider incremental measures to update incrementally as the corpus grows (Wen et al., 2002).

With respect to practical implications, the proposed approach could be used for rendering value-added services in various fields such as business and education (Liu et al., 2006; Adamic et al., 2008). From a business perspective, QA could be used as a collaborative medium for knowledge sharing in organizations. The proposed approach thus finds its importance in harnessing knowledge generated by diverse users such as customers or employees of an organization in answering newly posed questions.
In an educational environment, collaborative tools similar to a CQA service can be used to pervasively ask and answer subject related questions as a community. The proposed approach, if applied to such services, could automatically answer student queries related to a subject domain. The next step for Information Systems researchers is to take the lead in using the proposed approach to deploy new socio-technical systems.

References


Proceedings of the 28th International ACM Conference on Research and Development in Information Retrieval, Salvador, Brazil, 617-618.