Multidimensional Scaling Based Knowledge Provision for New Questions in Community Question Answering Systems

Siqi Xiang[1], Wenge Rong[1], Yikang Shen[2], Yuanxin Ouyang[1], Zhang Xiong[1]
[1]School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[2]Sino-French Engineer School, Beihang University, Beijing 100191, China
{sqxiang, w.rong, yikang.shen, oyyx, xiongz}@buaa.edu.cn

***Abstract*—Community-based Question Answering (CQA) sites have become popular since they allow users to get answers to complex, detailed and personal question from other users directly. However, since answering a question depends on the ability and willingness of other users to address the askers' real needs, a significant fraction of the questions remain unanswered. To decrease the unanswered question rate and then improve the user experience, in this paper, a multidimensional scaling (MDS) based data reorganization method is proposed. By using this method, the CQA system can predict the askers' intention and accordingly provide related previous question/answer pairs to help them find useful information. The method has been evaluated on an off-line dataset extracted from Baidu Zhidao and the result has shown its promising potential in knowledge management in CQA systems.**

***Keywords*-Community-based Question Answering; Knowledge Management; Multidimensional Scaling**

I. Introduction

Recently Community Question Answering (CQA) sites such as Yahoo! Answers[1], Baidu Zhidao[2] have become one of the most important platforms for users to obtain useful information. By using CQA systems, users can ask questions via sentences rather than issuing a query in form of keywords to a Web search engine. CQA has proven success for knowledge sharing since natural language is easier for users to express their real information needs [1]. Furthermore, using CQA is also easier to get answers of personal, extremely specific and/or open-ended questions as in these cases it is difficult for search engine to directly provide such complex and heterogeneous information [2].

Though CQA has shown its promising potential in information seeking, there are also a lot of challenges among which unanswered question is one of the most important one. It is frequently observed that in spite of active participation in CQA sites, a significant portion of questions remain unanswered. This phenomenon widely exists in CQA systems and is often referred as *question starvation* [3].

Several efforts have been made to reduce the amount of unanswered questions and these approaches can be roughly divided into three categories. One kind of approaches is going to pro-actively push open questions to the most relevant potential answerers to get prompt and qualified answers [4], [5], [6]. The other type of approaches attempts to automatically generate answers from external knowledge resources such as Wikipedia [7]. The last kind of methods tries to answer new questions by reusing past solved questions within the CQA site itself [8], [9]. Compared with the first two approaches, the last one can avoid answer delay in waiting experts' response and also alleviate the users efforts to search useful information in relatively large documents. As such it has been attached much importance within CQA's community.

The past experience based approaches rely on the intuition that even if personal and narrow, some questions are recurrent enough to allow for at least a few new questions to be answered by past material. Actually from the recently analysis, 25% of question in certain categories of Yahoo! Answers are recurrent, at least at the question-title level over a period of one year [8]. Though the idea of finding answer from past question/answer (Q/A) pairs is enlightening, previous works have also revealed several its own challenges.

One of the main difficulties for past Q/A based information seeking is the relevance detection between new questions and the Q/A pairs from the past. To solve this problem, a large number of methods have been proposed and the most widely used and straightforward one is the bag-of-words based methods, due to its simplicity and robust implementation [10]. However, this kind of approaches is difficult to evaluate the semantic or meaning similarity among questions and answers. As such some advanced methods are further proposed by capturing semantic [11], [12] and/or topic information [13] to support similarity calculation.

The similarity obtained by these methods is normally s-tatistic. However CQA sites are naturally dynamic as a large number of new questions/answers are accumulated everyday [14]. Furthermore, since in reality each question is personalized, which means that same intention can be explain in different ways [15]. The intention beyond the question is more important since queries are normally highly dynamic [16]. Inspired by this idea, in this paper, we proposed a novel mechanism to integrate the dynamics property into relevance identification process.

We deal with this issue by organizing past Q/A pairs

[1]http://answers.yahoo.com
[2]http://zhidao.baidu.com

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

similarity with a modified multidimensional scaling (MDS) [17], [18] method, thereby recovering the correlation between Q/A pairs and reconstructing the latent structure. The most relevant past Q/A pairs with the new question are then found according to distribution of past Q/A pairs in the space. Relevant Q/A pairs will be retrieved no matter whether they use similar words in new questions and irrelevant and improper information will be excluded. Experimental study conducted on dataset from Baidu Zhidao has shown its capability in meeting users query intention.

The rest of this paper is organized as follows. Background and related work will be investigated in Section 2. The proposed method will be illustrated in section 3 and section 4 will elaborate and discuss the experimental study. Finally section 5 will give conclusion and point out possible future research directions.

II. Background

A. CQA Based Information Seeking

To meet our requirement of efficiently and effectively extracting information on-line, a lot of advanced information communication technologies (ICTs) are proposed among which community based question/answering (CQA) forum has become one of the most influential mechanism. Unlike traditional search engines which have proven success during last decades, CQA support users to find and share knowledge with other users. Furthermore, different from search engine which employs keywords based search, in CQA based forums users can use natural language to communicate with each other.

Though the benefits of CQA based information sharing is well documented, there is also several challenges and one of them is to provide better user experience for those questions without prompt reply. One of the possible solutions is to provide past question and answer information with regard to the new question. This method is based on the finding that past questions and answers are recurrently to some extend for providing useful information to the new questions [8]. To better improve the user experience for CQA based information seeking, there is a key requirement of finding relevant question/answer (Q/A) pairs with regard to the new question.

To find the relevant information from past Q/A pairs, a key technology is to calculate the similarity between the new question and past Q/A materials. However, it is difficult to find a proper method to evaluate the similarity, particularly in CQA systems since in CQA systems sentences or paragraphs are normally short document, which has aggravated the weakness. A lot of researches have been devoted to find high quality similarity, e.g., topic-based [19], category-based [20], context-based [21], and etc.

After obtain the similarity, it still needs to consider more information before using it for past Q/A pairs recommendations. In fact, each question is personalized and has temporal popularity [16] and the questions can be explain in different ways. This means the calculated similarity is always different from the real similarity and calls for advanced mechanism to reveal this difference. In this paper, a novel multidimensional scaling based mechanism is proposed to support the similarity calibration.

B. Multidimensional Scaling

Multidimensional scaling (MDS) [17] refers to the general task of assigning Euclidean coordinates to a set of objects such that given a set of dissimilarity, similarity, or ordinal relations between the objects, the relations are obeyed as closely as possible by the embedded points. This assignment of coordinates is also known as a Euclidean embedding.

MDS algorithms fall into two broad classes: metric algorithms, which seek an embedding with inter-point distances closely matching the input dissimilarities; and non-metric algorithms, which find an embedding respecting only the relative ordering of the input dissimilarities. In this paper we propose a hybrid method by using mixed metric and non-metric algorithm to solve similarity calibration problem.

Reasons of using MDS model can be conclude as:

1) MDS has shown its capacity for data clustering. Because entities are placed into a MDS space with respect to their proximities, entities with strong relation will be naturally clustered.
2) Profit from restriction between distance. MDS model can also find and smooth out noise produced by inaccurate measure method. Even facing incomplete proximity matrices, it is also robust.
3) MDS has been proven useful in semantic analysis [22].

C. Problem Definition

In previous sections we illustrated the motivation and importance of similarity calibration for CQA based information seeking system. After briefly introducing the inspiration of employing MDS to implement similarity estimation and tuning, here the targeted problem is formally defined as a two-stage process.

In a CQA site, normally a question and answers are presented in a web page. For a given page i, t_i is defined as the information which contains the question and answers of that page i. Therefore, similarity between any two pages i and j is defined as s_{ij} and calculated between any pair of text t_i and t_j.

As introduced before, this paper's goal is to find most relevant texts to a new question, whether these texts contains words used in the question is not a necessary condition. In order to realize this goal, two problems are realized:

1) Mapping texts from Q/A history into an m dimension space as perfect points using MDS technique. In this space the Euclidean distance represent proximity between points. If a pair of point has a strong relation, their distance d_{ij} should be short, and vice versa.
2) Present the new question in the same space by one or several points, in order to find related past Q/A pairs close to these points.

It bears pointing out that targeted problem is different from currently existing similar question recommend system. Exiting

works focus on using keyword or semantics oriented search methods to find similar question, while in this paper we tried to pre-organize data from Q/A history by using MDS technique to recover latent structure. With the history data increasing gradually, the similarity between any different pairs will be changed as new Q/A pairs are inserted into the space. It is believed that this idea can reflect the intuition of maintaining dynamic similarity among different Q/A pairs.

III. MDS based Q/A Pairs Identification

A. Pre-Processing

MDS model require only rough estimation of similarity between any two pages and estimation can be optimized gradually by MDS itself. In order to calculate similarity between texts in CQA forum, each page was firstly transformed into a set of words u_i. The similarity between any two pages are defined by using Jaccard similarity approach [23] and it is defined as below:

$$p_{ij} = \frac{|u \bigcap v|}{|u \bigcup v|} \tag{1}$$

After obtain similarities, we need to map these similarities into a dissimilarity matrix D. Because in our MDS space, distance between pairs reflect the dissimilarity between pages. Therefore in this paper we define dissimilarity as below:

$$\delta_{ij} = \frac{b}{p_{ij}}, b = 10 \tag{2}$$

where b is a scaling constant and has no influence on final result.

B. Hypothesis

It is intuitive to think that the real relevance of new different texts is always influenced by some noise. As such the similarity value is a combination of real proximity and inevitable noise, as defined below:

$$p_{ij} = \rho_{ij} + \epsilon_{ij} \tag{3}$$

where ρ_{ij} refers to real similarity, ϵ_{ij} refers to noise(or error).

Here we propose a hypothesis and argue that a similarity can be accepted only the noise or error is within a certain bound E, i.e., $\left|\epsilon_{ij}\right| < E$. Thus if $p_{ij} \gg E(p_{ij} > 10E)$, which means the noise is less influential, then p_{ij} is reliable and need to be taken into account. If $10E > p_{ij} > E$, the situation is critical, the interval $[E, 10E]$ is then called the critical interval. If $p_{ij} \approx E$ or $p_{ij} < E$, which means the noise is large enough, as a result p_{ij} is worthless and need to be rejected.

In practice, we can choose a value in critical interval as bound value B_p, reject all $p_{ij} < B_p$. The choice of B_p is empirical and will be discussed in the section of experimental study. The transform function from p_{ij} to δ_{ij} can be redefine as

$$\delta_{ij} = \begin{cases} \frac{b}{p_{ij}} & , \quad p_{ij} > B_p \\ \infty & , \quad p_{ij} < B_p \end{cases}, b = 10 \tag{4}$$

From Eq. 4, we can obtain the dissimilarity matrix Δ. With Δ we can still find subsets between which only ∞ dissimilarity exist. In fact, this is reasonable, because these subsets is about different subject. In practice, for relations satisfy $p_{ij} > B_p$, we treat them as a metric problem. For relations satisfy $p_{ij} < B_p$, we treat them as non-metric problem. The only constraint is that their dissimilarity δ_{ij} should bigger than most of the dissimilarity δ_{ij} which is not ∞ in Eq. 4. After these steps, we can build a metric and non-metric mixed MDS.

C. Latent Structure

In order to construct and describe a MDS space. We need the following basic definitions.

- D1 n denotes the number of pages.
- D2 If a similarity is smaller than bound value for a pair of page, i and j, a dissimilarity value δ_{ij} is given. If δ_{ij} is ∞, we speak of a missing value, because the correct similarity is covered by noise.
- D3 A dissimilarity is a proximity that indicates how dissimilar two objects are. A small score indicates that the objects are similar, a high score that they are dissimilar.
- D4 X denotes (a) a point configuration (i.e., a set of n points in m-dimensional space) and (b) the $n \times m$ matrix of the coordinates of the n points relative to m Cartesian coordinate axes. A Cartesian coordinate system is a set of pairwise perpendicular straight lines (coordinate axes). All axes intersect at one point, the origin, O. The coordinate of a point on axis a is the directed (signed) distance of the points perpendicular projection onto axis a from the origin. The m-tuple $(x_{i1}, ..., x_{im})$ denotes the coordinates of point i with respect to axes $a = 1, ..., m$. The origin has the coordinates $(0, ..., 0)$.
- D5 The Euclidean distance between any two points i and j in X is the length of a straight line connecting points i and j in X. It is computed by the value resulting from the formula

$$d_{ij} = \left[\sum_{a=1}^{m} \left(x_{ia} - x_{ja}\right)^2\right]^{1/2} \tag{5}$$

where x_{ia} is the coordinate of point i relative to axis a of the Cartesian coordinate system. We also use $d_{ij}(X)$ for the distance to show explicitly that the distance is a function of the coordinates X.

So far, the task of MDS was defined as finding a low-dimensional configuration of points representation objects such that the distance between any two points matches their dissimilarity as closely as possible. Of course, we would prefer that each dissimilarity should be mapped exactly into its corresponding distance in the MDS space. But that requires too much, because empirical data always contain some component of error. Traditionally error of representation is defined as

$$e_{ij}^2 = \left(d_{ij} - \delta_{ij}\right)^2 \tag{6}$$

But Eq. 6 cannot be apply directly into our situation. In our situation, not all δ_{ij} is defined. We still need a define, if $\delta_{ij} = \infty$

$$e_{ij}^2 = \left(\frac{2b/B_p}{d_{ij}}\right)^2 \tag{7}$$

From Eq. 7 we want to explain that similarity between page i and j is a miss value, but points i and j cannot be very close, because their similarity should be smaller than $B_p + E$

$$\rho_{ij} < B_p + E \tag{8}$$

Summing over i and j yields the total error (of approximation) of an MDS representation

$$\sigma_r = \sum_{i<j} e_{ij}^2 \tag{9}$$

Thus, our aim is to find a coordinate matrix X to make $\sigma_r(X)$ minimal.

D. Minimization of $\sigma_r(X)$

In order to minimize $\sigma_r(X)$, we choose gradient descent algorithm. But with the augmentation of pages in dataset, it become harder and harder to minimize $\sigma_r(X)$, and the accuracy of coordinate decrease rapidly. In order to avoid these trouble and facilitate calculate. We choose to minimize one point's error each time. Minimization algorithm can be summarized as shown in Fig. 1.

Input: X, k
1: $k = 0$
2: set $flag = True$
3: **while** $flag = True$ and $k \leq$ maximum iteration **do**
4: $k++$
5: Minimize each point's error
6: $\sigma_i(x_{i1}, ..., x_{im}) = \sum_j e_{ij}^2$
7: **if** No point found a new position **then**
8: $flag = False$
9: **end if**
10: **end while**

Fig. 1. Minimization algorithm

Meanwhile, we have mentioned that $\exists \Delta_i', \Delta_j' \subset \Delta\left(\Delta_i' \cap \Delta_j' = \emptyset\right)$ between which only ∞ dissimilarity exist. Which means that we can separate Δ into different subsets Δ_i', and find coordinate matrix X_i separately. In practice, points is been added to X one by one, as shown in Fig. 2. In this way we can easily extend dataset.

Input: X, i
1: $i = 1$
2: **while** $i < n$ **do**
3: $i++$
4: Set a random coordinate to pages i, add this coordinate as a new line to X
5: Minimize $\sigma_r(X)$
6: **end while**

Fig. 2. Adding new point

E. Dimensionality

So far, we have not discuss about the dimension m. But dimensionality is an important parameter in MDS. Scaling with too few dimensions may distort the true and reliable MDS structure due to over-compression or may lead to technical problems. On the other hand being too generous on dimensions may blur the MDS structure due to over-fitting noise components. In ordinal MDS, any matrix of dissimilarity $\delta_{ij}(i < j)$ can be represented, with zero total error, in $m = n \times 2$ dimensions.

We plot the resulting total error values (on the Y-axis) against the m-values (on the X-axis, then looks for an elbow in this curve, a point where the decrements in Stress begin to be less pronounced. The rationale of this choice is that the elbow marks the point where MDS uses additional dimensions to essentially only scale the noise in the data, after having succeeded in representing the systematic structure in the given dimensionality m.

F. Mapping new question into MDS Space

After MDS space is constructed, the next step is to add the new questions into the same question and then find related Q/A pairs for knowledge sharing. The process is same as adding a Q/A pairs into the MDS space and then return nearby pages as the mapping result.

In this paper, we express new question's error of fitness by $\sigma_k(x_1, ..., x_m)$, which is a function of position. We start with a random position and use gradient descent algorithm to find minimal point. But σ_k has several local minimals, these local minimals are not useless like in other minimization problem. Because each local minimals can represent a user's potential search intention. In practice we set pages around each local minimal as a result subset. These subset are shown to user, and ranked by their error value σ_k.

IV. Experimental Study

A. Data Collection

For the purpose of evaluating and validating the proposed MDS based similarity calibration model, a dataset is installed by collecting data from Baidu Zhidao, which is the biggest Chinese CQA site. We randomly collected 146,063 questions during the period of March 2013 from this CQA site. Among these questions, 101,316 (69.3%) questions have at least one answer, 66,638 (45.6%) questions have been solved, while 42,747 (30.7%) questions do not receive any answer after 6

months after it is posted online. This finding is accordance of previously report [3].

B. Baseline

We compare our MDS approach against four popular retrieval models:
1) TF-IDF,

$$S_{q,d} = \frac{\sum_{i=1}^{N} w_{q,t} w_{d,t}}{\sqrt{\sum_{i=1}^{N} w_{q,t}^2} \sqrt{\sum_{i=1}^{N} w_{d,t}^2}} \tag{10}$$

$$w_{q,t} = ln(1 + \frac{N}{f_t}), w_{d,t} = 1 + ln(tf_{t,d}) \tag{11}$$

Here N is the number of questions in the whole collection, f_t is the number of questions containing the term t, and $tf_{t,d}$ is the frequency of term t in d.
2) Okapi model (Okapi) [24],

$$\begin{aligned} &S_{q,d} = \\ &\sum_{i=1}^{n} \mathrm{IDF}(q_i) \cdot \frac{f(q_i, d) \cdot (k_1 + 1)}{f(q_i, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{\mathrm{avgdl}})} \\ &\mathrm{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \end{aligned} \tag{12}$$

3) Language model (LM) [25],

$$\begin{aligned} &S_{q,d} = \prod_{t \in q} ((1 - \lambda) P_{ml}(t|d) + \lambda P_{ml}(t|Coll)) \\ &P_{ml}(t|d) = \frac{tf_{t,d}}{\sum_{t' \in d} tf_{t',d}} \\ &P_{ml}(t|Coll) = \frac{tf_{t,Coll}}{\sum_{t' \in Coll} tf_{t',d}} \end{aligned} \tag{13}$$

4) Translation model (TM) [26][27],

$$\begin{aligned} S_{q,d} = \prod_{t \in q} ((1 - \lambda) \sum_{w \in d} T(t|w) P_{ml}(w|d) \\ + \lambda P_{ml}(t|Coll)) \end{aligned} \tag{14}$$

The model parameters setup are same as in [28].

C. Evaluation Metrics

To evaluate the methods, inspired by the idea of $P@K$ in the domain of information retrieval, which evaluate if the proper pages appear in the top K return result. We use metrics called $S@N$ and $R@N$, which are similar to [29], to indicate whether within top N returned Q/A pairs the users can solve their question ($S@N$) or find related information ($R@N$). As to the selection of the value of N, since a small one will have not real impact on the test result, as indicated in [29], while a bigger one will cost users too much effort to study, in this experiment it is set to 5, which is typically used in the real applications. For example, in Baidu Zhidao, when a user asks a question, normally 5 more questions are listed below for the user's reference. A question is marked solved if user think he/she has obtain enough information to solve the question or a way of solving the question is indicated. A question is marked related if user get some useful information, but an answer is not yet obtained.

We also evaluate the performance of our approaches using Mean Average Precision (MAP) [30], Mean Reciprocal Rank (MRR) [31]. MAP rewards approaches which return relevant questions early and also rewards correct ranking of the results. MRR gives us an idea of how far down in a ranked list we must look to find a relevant question. In process of experimental study, we found that some of unanswered questions are not able to be answered, because these "questions" are either declarative sentence or do not make any sense. In fact, this kind of non-sense question is an important issue in CQA data analysis [32].

D. MDS Space Construction

As elaborated in previous section, the most important part of the proposed approach is to construct a MDS space for the Q/A pairs from the past. In this section, the MDS space construction will be illustrated with a small subset of the data collected from Baidu Zhidao to present its process during which the parameters definition will be also presented. To simplify the process and present the MDS space in an understandable perspective, we only use 250 Q/A pairs in this step while it is the same process with the total number of dataset.

The first parameter which must be defined is the noise or error bound, E. After analyze the dataset, the similarity distribution is obtained and presented as in Fig. 3. It is observed that most of the similarities between Q/A pairs is lower than 0.015. Since most pages in this dataset is about different subject, as such 0.015 can be taken as the value of E.

After realization of error threshold E, the next step is to get the critical interval bound value B_p. Before mapping Q/A pairs into the MDS space, it is necessary to calculate the dissimilarity between these Q/A pairs. For convenience, in the rest of paper, we use bound value of dissimilarity $B_d = \delta / B_p$ instead of B_p, where δ is the dissimilarity defined in Eq. 2.

After observation of dataset, it is found that the strong relations exist mostly between a limited number of texts and the dissimilarity value is within 50 and 100. Therefore, to empirically identify the interval bound value, we have build different MDS space with different B_d value of 50, 60, 70, 80, 90, 100, respectively. To simplify the presentation, the values are projected into a three dimension space and Figs. 4, 5, and 6 are the spaces when B_d equals 100, 80, 60, respectively[3].

[3]In Figs. 4, 5, and 6 each entry is represented by a entry number. In the rest of the paper, Q/A pairs is often represented by their page number, which is actually the last part of there URL. For Baidu Zhidao, it's *http://zhidao.baidu.com/question/ + entry number*.

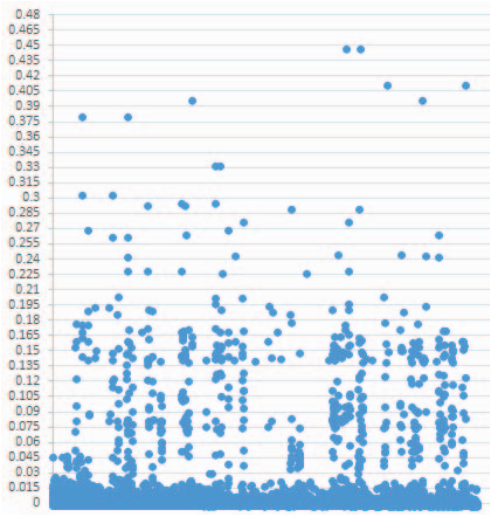

Fig. 3. Similarities between 250 Q/A pairs

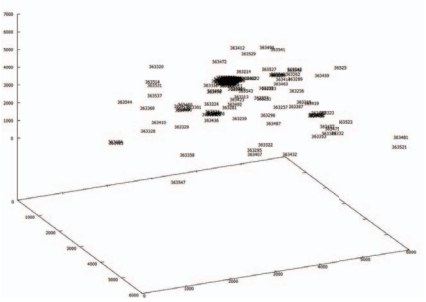

Fig. 4. MDS Configuration while $B_d = 100$

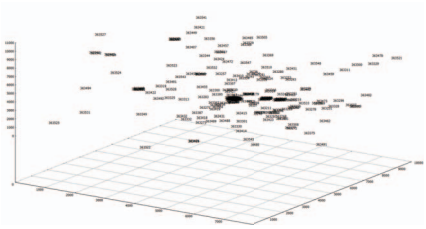

Fig. 5. MDS Configuration while $B_d = 80$

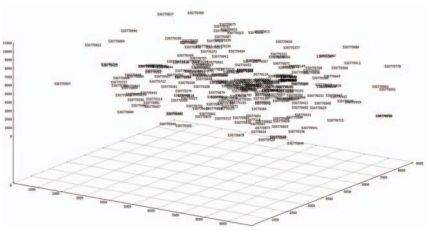

Fig. 6. MDS Configuration while $B_d = 60$

After analysis of those figures in term of different B_d, it is observed that a high density area exists at the center of space, and when zooming in this area further, entries of different subjects can be founded. In fact, this problem is raised by a big value of B_d and accordingly the 'indirect attraction' between points.

Indirect Attraction: Assuming that dissimilarity δ_{ij} between page i and j is bigger than B_d. But there exist several pages ks that δ_{ik} and δ_{kj} is both smaller than B_d. Thus distance d_{ij} has a great chance to be smaller than B_d.

According to 'indirect attraction', it is able to explain the 'black hole' at the centre of space: the bigger B_d is, the more points are attracted to the 'black hole'. In order to avoid appearance of 'black hole' we need to decrease B_d to a reasonable value.

Fig. 6 shows a more reasonable configuration. In this space, Q/A pairs with similar subject are gathered together. For example, Table I shows some of questions that has been gathered in this space. We can see that all entries are about Samsung consumer electronics products. Although, their are clearly two kind of questions, i.e., hardware problem and software problem. When we further look into the answers of these questions, it is found that all of these questions are answered by a user named 'Samsung digital service platform' who is the Samsung customer service. Therefore there is a strong tie between these four Q/A pairs. As such one can easily draw a conclusion from this Q/A pair group that 'Samsung digital service platform' can solve problems related to Samsung devices.

TABLE I
TITLE OF QUESTIONS IN FIG. 5

Entry No.	Entry Title
530770098	My old Samsung mobile phone is broken, what should I do?
530770121	I need to repair my Samsung tablet, where should I go?
530770501	Can I play rmvb video on my Samsung cellphone?
530770805	How can I change my Samsung mobile phone ringtone?

TABLE II
DISSIMILARITY AND DISTANCE BETWEEN Q/A PAIRS IN TABLE I

Entry A	Entry B	d_{ij}	δ_{ij}
530770098	530770121	51.17591514	51.15384615
530770098	530770501	68.42034818	68.51301115
530770098	530770805	65.23242261	65.14285714
530770121	530770501	118.0378992	77.46022216
530770121	530770805	109.1490554	78.69822485
530770501	530770805	52.23240898	52.25404026

Table II shows the positive aspect of 'Indirect Attraction'. The dissimilarity between some pairs is higher than bound value, but their distance is smaller than bound value. We came to conclusion that 'Indirect Attraction' can help us to recover missing value.

This observation make us set 60 as a empirical bound value of dissimilarity δ_{ij}. Corresponding similarity bound value $B_p = b/60 = 0.133$

$$E < B_p = 0.133 < 10E \tag{15}$$

It is found that Eq. 15 satisfied the constraint condition of B_p.

E. Knowledge for New Questions

In previous section we have illustrated the steps to empirically realize the parameters for MDS space. In this part, we will use the whole dataset to evaluate the MDS based knowledge management for new questions in CQA sites. Particularly, we use the solved Q/A pairs to try to provide information for the unsolved questions in the dataset.

In Figs. 4 to 6 the dimension of MDS space is three to make it understandable for parameter illustration. However, the dimension can be set to a larger one to make full use of MDS advantages. To find the proper dimension, it is necessary to analyse the dataset.

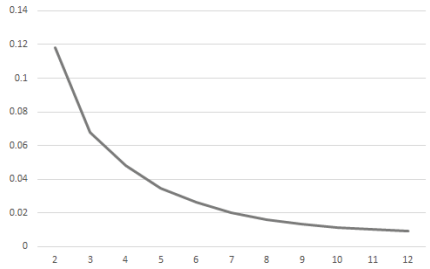

Fig. 7. Evolution of average stress σ_r/n

Fig. 7 shows the evolution of average stress and the Y-axis is proportional to average stress σ_r/n. X-axis is dimension. It is found that after dimension increase to 10 the decrements in Stress begin to be less pronounced. As a result in the experimental study dimension $m = 10$ is selected.

TABLE III
EXPERIMENTAL STUDY COMPARISION

	$S@5$	$R@5$	MAP	MRR
MDS	**4%**	29.4%	**0.408**	**0.631**
TM	3.8%	**31.4%**	0.401	0.605
LM	3.1%	25.8%	0.381	0.583
Okapi	3.0%	25.1%	0.339	0.542
TF-IDF	2.2%	24.07%	0.243	0.451

Table III show the rate of solved and related question of the proposed MDS based method against traditional TF-IDF based method. From the analysis it is found that about 4.0% of question has been solved and 29.4% questions are marked as been given useful related information. This experiment just shows the ability of how past answered question can be used to help solved unanswered question, and the possibility to use them automatically.

Comparing with TF-IDF, Okapi and LM, MDS has a significant augment on all metrics. In reality, a common observation is that tradition information retrieval method, like TF-IDF, tend to give relevant pages at the top of search results, the rest of results are useless. Thus, while our system extend these top search results using MDS space, we actually extend result with more relevant information and context. These extra information give user a better chance to solve their problem. The most interesting thing is that these extra relevant Q/A pairs don't need to contain same word used in new question, which is the main advantage of MDS method.

Comparing with TM model, MDS is better in resolved rate and MRR, close in MAP, but worse in related rate. Therefore, we can see that MDS tend to be more accuracy at top search results, and keep the related rate and MAP at a high level. Both of this two methods have ability to retrieve relevant Q/A pairs using different word. TM model achieve this goal through using the co-exist probability of different words between paraphrase, which give LM model a better chance to find related question. MDS space extend these top search results with relevant information and context, and it also re-rank the search result base the inter-relationship between these question. Thus the rank order is more accurate.

V. Conclusion and Future Work

In this paper, we explored the possibility of reduce low response rate problem through MDS model. Two major problems have been studied, i.e., (i) construction of MDS space for dataset of texts, (ii) finding relevant past Q/A pairs for a new question. Experiment result confirmed our idea, showing that this method can give valuable information to help people deal with their question. Finally, our approach of analyse data through MDS space can be also applied to other domain, e.g., social network data, financial data. The only pre-requirement is the proximity should be available.

Our experiments were carried on only one source of Q/A archive. Our future work will test the proposed MDS model on FAQ (Frequently Asked Questions) archives. The idea of compiling knowledge into FAQ files has existed for some time and has been proved to be a useful supplement for Q/A based information seeking [33]. We can utilize FAQ files to retrieve answers efficiently by answer reuse, and to find possible answer that may not be found by narrowly scoped question answering methods. We also plan to do experiments on data from Yahoo!Answers, which is a much larger collection in English questions than Baidu Zhidao.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (No. 61332018), the National Department Public Benefit Research Foundation of China (No. 201510209), and the Fundamental Research Funds for the Central Universities.

References

[1] M. W. Bilotti, J. L. Elsas, J. G. Carbonell, and E. Nyberg, "Rank learning for factoid question answering with linguistic and semantic constraints," in *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, 2010, pp. 459–468.

[2] A. Y. K. Chua and S. Banerjee, "So fast so good: An analysis of answer quality and answer speed in community question-answering sites," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 10, pp. 2058–2068, 2013.

[3] B. Li and I. King, "Routing questions to appropriate answerers in community question answering services," in *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, 2010, pp. 1585–1588.

[4] D. Horowitz and S. D. Kamvar, "The anatomy of a large-scale social search engine," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 431–440.

[5] G. Dror, Y. Koren, Y. Maarek, and I. Szpektor, "I want to answer; who has a question?: Yahoo! answers recommender system," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1109–1117.

[6] Z. Zhao, F. Wei, M. Zhou, and W. S. H. Ng, "Cold-start expert finding in community question answering via graph regularization," in *Proceedings of the 20th International Conference on Database Systems for Advanced Applications*, 2015.

[7] E. M. Rodrigues and N. Milic-Frayling, "Socializing or knowledge sharing?: characterizing social intent in community question answering," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009, pp. 1127–1136.

[8] A. Shtok, G. Dror, Y. Maarek, and I. Szpektor, "Learning from the past: answering new questions with past answers," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 759–768.

[9] H. Wu, W. Wu, M. Zhou, E. Chen, L. Duan, and H.-Y. Shum, "Improving search relevance for short queries in community question answering," in *Proceedings of the 7th ACM international conference on Web search and data mining*, 2014, pp. 43–52.

[10] M. W. Bilotti, P. Ogilvie, J. Callan, and E. Nyberg, "Structured retrieval for question answering," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 351–358.

[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31th International Conference on Machine Learning*, 2014, pp. 1188–1196.

[12] W.-t. Yih, X. He, and C. Meek, "Semantic parsing for single-relation question answering," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 643–648.

[13] Y. Wu, W. Wu, Z. Li, and M. Zhou, "Mining query subtopics from questions in community question answering," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

[14] P. Fichman, "A comparative assessment of answer quality on four question answering sites," *Journal of Information Science*, vol. 37, no. 5, pp. 476–486, 2011.

[15] J. Hu, G. Wang, F. Lochovsky, J.-t. Sun, and Z. Chen, "Understanding user's query intent with wikipedia," in *Proceedings of the 18th International Conference on World Wide Web*, 2009, pp. 471–480.

[16] A. Kulkarni, J. Teevan, K. M. Svore, and S. T. Dumais, "Understanding temporal query dynamics," in *Proceedings of the 4th International Conference on Web Search and Web Data Mining*, 2011, pp. 167–176.

[17] I. Borg, *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[18] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. J. Kriegman, and S. Belongie, "Generalized non-metric multidimensional scaling," in *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, 2007, pp. 11–18.

[19] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, "Searching questions by identifying question topic and question focus," in *Proceedings of the 46th Annual Meeting of the Association for Computational LinguisticsCL*, 2008, pp. 156–164.

[20] X. Cao, G. Cong, B. Cui, and C. S. Jensen, "A generalized framework of exploring category information for question retrieval in community question answer archives," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 201–210.

[21] G. Zhou, L. Cai, J. Zhao, and K. Liu, "Phrase-based translation model for question retrieval in community question answer archives," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011, pp. 653–662.

[22] E. Cambria, Y. Song, H. Wang, and N. Howard, "Semantic multidimensional scaling for open-domain sentiment analysis," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 44–51, 2014.

[23] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 1.

[24] J. Jeon, W. B. Croft, and J. H. Lee, "Finding similar questions in large question and answer archives," in *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, 2005, pp. 84–90.

[25] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, "Searching questions by identifying question topic and question focus," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 2008, pp. 156–164.

[26] X. Xue, J. Jeon, and W. B. Croft, "Retrieval models for question and answer archives," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 475–482.

[27] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, and Y. Liu, "Statistical machine translation for query expansion in answer retrieval," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007.

[28] X. Cao, G. Cong, B. Cui, and C. S. Jensen, "A generalized framework of exploring category information for question retrieval in community question answer archives," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 201–210.

[29] M. Liu, Y. Liu, and Q. Yang, "Predicting best answerers for new questions in community question answering," in *Proceedings of 11th International Conference on Web-Age Information Management*, 2010, pp. 127–138.

[30] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 271–278.

[31] N. Craswell, "Mean reciprocal rank," in *Encyclopedia of Database Systems*. Springer, 2009, pp. 1703–1703.

[32] B. Li, T. Jin, M. R. Lyu, I. King, and B. Mak, "Analyzing and predicting question quality in community question answering services," in *Proceedings of the 21st International Conference Companion on World Wide Web*, 2012, pp. 775–782.

[33] V. Jijkoun and M. de Rijke, "Retrieving answers from frequently asked questions pages on the web," in *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, 2005, pp. 76–83.