Ranking Community Answers via Analogical Reasoning*

Xudong Tu‡, Xin-Jing Wang†, Dan Feng‡, Lei Zhang†
†Microsoft Research Asia. {xjwang, leizhang}@microsoft.com
‡HuaZhong Univ. of Sci. and Tech., {tuxudong, dfeng}@{smail.}hust.edu.cn

ABSTRACT
Due to the lexical gap between questions and answers, automatically detecting right answers becomes very challenging for community question-answering sites. In this paper, we propose an analogical reasoning-based method. It treats questions and answers as relational data and ranks an answer by measuring the analogy of its link to a query with the links embedded in previous relevant knowledge; the answer that links in the most analogous way to the new question is assumed to be the best answer. We based our experiments on 29.8 million Yahoo!Answer question-answer threads and showed the effectiveness of the approach.

Categories and Subject Descriptors
G.3 [**Probability and Statistics**]: *correlation and regression analysis.* H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *web-based services.*

General Terms
Algorithms, Experimentation.

Keywords
Community Question Answering, Analogical Reasoning

1. INTRODUCTION
As one type of portals for fast-growing *user generated content (UGC)*, community question-answering (CQA) sites have attracted a large number of users both seeking and providing answers to a variety of questions on a variety of subjects. Since they allow anyone to post or answer any questions on any subjects, the quality of answers varies greatly. Thus, the ability to automatically identify the best answers has significant impact on users' satisfaction.

The major challenge of identifying high-quality answers in CQA sites is the lexical gap between a question and its correct answer. The gap is caused by at least two factors: 1) textual mismatch between questions and answers. Words appeared in a question will not necessarily reappear in its best answer; and 2) user generated spam or flippant answers.

Previous work solved this problem by either generating complementary features provided by highly structured CQA sites [1], or finding textual clues using machine-learning techniques [3]. In this paper, we propose an analogical reasoning-based ranking technique which uses previous relevant knowledge to bridge the lexical gap, and discovers non-textual clues between questions and their answers to identify the correct ones. In particular, we assume that a question has implicit positive links with its right answers, and negative linkages with other answers. Given a new question, firstly we retrieve a set of positively linked (i.e. correctly answered) question-answer pairs (q-a pairs) whose questions are similar to the new question from a knowledge base. The retrieved q-a pairs construct a supporting set. Secondly, we predict the analogy of the link between the new question and each of its candidate answers to the links in the supporting set. The analogy is predicted by a logistic regression model learnt offline. The answer which has the most analogous linkage is assumed to be the best answer to the new question.

There are two key aspects of the proposed idea: 1) Instead of either separating questions and their answers as independent information sources or mixing them as one object, we treat them as relational data and predict the property of the link. Secondly, instead of directly predict the link with the link prediction model, we use a supporting set to bridge the lexical gap and facilitate the prediction. Intuitively, not only the supporting set is more likely to share terms with the candidate answers, but also it provides the knowledge of how similar questions are answered, while the "way of answering" suggests the desired answers.

2. THE APPROACH
2.1 Learning the Link Prediction Model
We adopt the Bayesian Analogical Reasoning (BAR) framework proposed by Silva et al. [3] to predict the latent q-a linkages.

Formally, let $X^{ij} = [\Phi_1(Q^i, A^j), \Phi_2(Q^i, A^j), \dots, \Phi_K(Q^i, A^j)]$ be a K-dimensional feature vector of the q-a pair of question Q^i and answer A^j, where Φ defines the mapping $\Phi: \mathcal{Q} \times \mathcal{A} \to \mathcal{R}^K$. Let $C^{ij} \in \{0,1\}$ be an indicator of the type of the link between Q^i and A^j, where $C^{ij} = 1$ indicates a positive linkage and $C^{ij} = 0$ otherwise. Let $\Theta = [\theta_1, \theta_2, \dots, \theta_K]$ be the parameter vector of the logistic regression model to be learnt, we have

$$P(C^{ij} = 1 | X^{ij}, \Theta) = \frac{1}{1 + \exp(\Theta^T X^{ij})} \quad (1)$$

where $X^{ij} \in X_{train} = \{X^{ij}, 1 \le i \le D_q, 1 \le j \le D_a\}$ is a training q-a pair and D_q, D_a are the numbers of training questions and answers respectively. Generally we have $D_a \gg D_q$.

We use both positive q-a pairs (i.e. good answer) and negative q-a pairs (i.e. noisy answers) to train this model, while the effect of negative ones is embedded in a Gaussian prior.

To ensure the preciseness of the prior, firstly we fit a BAR model using Maximum Likelihood Estimation (MLE) on the training data, and obtain an initial $\widehat{\Theta}$. Then we compute the covariance matrix $\hat{\Sigma}$ as a smoothed version of the MLE estimated covariance:

$$(\hat{\Sigma})^{-1} = \frac{c}{N} \cdot (\boldsymbol{X}^T \widehat{\boldsymbol{W}} \boldsymbol{X}) \quad (2)$$

where c is a predefined number, N is the size of the training dataset, $\boldsymbol{X}$ is the $N \times K$ feature matrix of the training q-a pairs, either positive or negative. $\widehat{\boldsymbol{W}}$ is a diagonal matrix with $\widehat{\boldsymbol{W}}_{\boldsymbol{ii}} = \hat{p}(i) \cdot (1 - \hat{p}(i))$, and $\hat{p}(i)$ is the predicted probability of a positive link for the ith row of $\boldsymbol{X}$. The prior of Θ is then the Gaussian $\mathcal{N}(\widehat{\Theta}, \hat{\Sigma})$.

*The work was done during the first author's internship in MSRA.

WWW 2009, April 20-24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

We randomly sample a similar number of negative points as a positive population to balance the number of positive and negative training data.

2.2 Retrieving Relevant Previous Knowledge

In the testing stage, given a new q-a thread, we retrieve some relevant *positive* q-a pairs from 29.8 million q-a threads crawled from the Yahoo!Answers site.

We adopt traditional information retrieval techniques to find the supporting q-a set. In particular, let Q_q be a query question and its answer list be $\{A_q^i, 1 \le i \le M\}$, we retrieve those positive q-a pairs $\boldsymbol{S} = \{Q^1{:}A^1, Q^2{:}A^2, \dots, Q^L{:}A^L\}$ with high cosine similarity above a threshold:

$$\boldsymbol{S} = \{Q^i{:}A^i \,|\, \cos(Q_q, Q^i) > \lambda, i \in \{1, \dots, D\}\} \quad (3)$$

where D is the size of the crawled Yahoo!Answers database and λ is a threshold. Each question is represented in the bag-of-word model. The effect of λ is shown in Figure 2.

2.3 Answer Ranking by Analogical Reasoning

Given $\boldsymbol{S}$, we score the new q-a pairs $\{Q_q{:}A_q^1, Q_q{:}A_q^2, \dots, Q_q{:}A_q^M\}$ with Eq.(4) by measuring a marginal:

$$score(Q_q, A_q^j) = logP(C_q^j = 1 | X_q^j, \boldsymbol{S}, \boldsymbol{C^S} = 1) - logP(C^j = 1 | X_q^j) \quad (4)$$

where A_q^j is the j-th answer of the query question. X_q^j represents the features of (Q_q, A_q^j). $\boldsymbol{C^S}$ is the vector of link indicators for $\boldsymbol{S}$, and $\boldsymbol{C^S} = 1$ indicates that all the q-a pairs in $\boldsymbol{S}$ have positive links, i.e. $\{C^1 = 1, C^2 = 1, \dots, C^L = 1\}$.

The first term in Eq.(4), i.e. $logP(C_q^j = 1 | X_q^j, \boldsymbol{S}, \boldsymbol{C^S} = 1)$, measures the probability of a positive query linkage when the supporting set is observed. The second term, $logP(C^j = 1 | X_q^j)$, on the other hand, evaluates this probability in the case that only the query q-a pair is observed. The idea underlying Eq.(4) is actually to measure to what extent (Q_q, A_q^j) would "fit into" $\boldsymbol{S}$, or to what extent $\boldsymbol{S}$ *explains* (Q_q, A_q^j). The more analogous it is to the supporting linked q-a pairs, the higher score a q-a pair obtains.

We use the learnt logistic regression model to predict the probability of a positive linkage in particular, which gives Eq.(5) and Eq.(6):

$$P(C_q^j = 1 | X_q^j, \boldsymbol{S}, \boldsymbol{C^S} = 1) = \int P(C^j = 1 | X_q^j, \Theta) P(\Theta | \boldsymbol{S}, \boldsymbol{C^S} = 1) d\Theta \quad (5)$$

$$P(C_q^j = 1 | X_q^j) = \int P(C^j = 1 | X_q^j, \Theta) P(\Theta) d\Theta \quad (6)$$

3. EVALUATION

We crawled 29.8 million questions from the Yahoo!Answers website, all of which have user-labeled "best answers". 100,000 randomly selected q-a pairs were used to train the link prediction model, and 16,000 q-a threads were used for testing; each contains 12.35 answers on average, which results in about 200,000 testing q-a pairs.

We compared our approach with three baseline methods: Nearest Neighbor (NN), Cosine distance and Bsets [2]. The first two directly measure the similarity between a question and an answer,

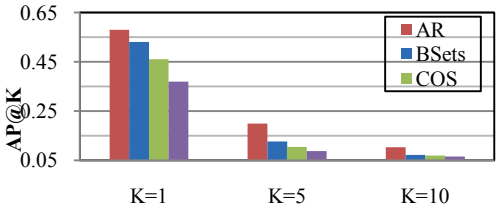

Figure 1. Average precision of our method and the baselines.

Table 1. MRR for our method and the baselines

Method	MRR	Method	MRR
NN	0.56	BSets [2]	0.67
Cosine	0.59	Our method	0.78

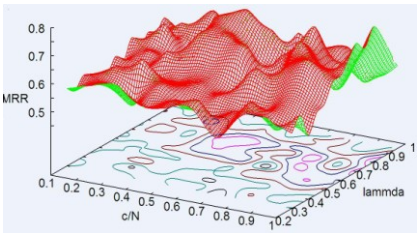

Figure 2. The Effect of similarity threshold λ and Prior scalar $\frac{c}{N}$. $\lambda = 0.8, \frac{c}{N} = 0.6$ gives best performance.

while the Bsets method leverages a supporting set to measure the similarity, however it does not treat questions and answers as relational data.

Figure 1 illustrates the mean average precision of the four methods on top $K = 1, 5, 10$ results, and Table 1 gives the Mean Reciprocal Rank (MRR) performance. Both show that our method significantly out-performed the baseline methods.

Figure 2 illustrates the effect of parameters $\frac{c}{N}$, the scalar in Eq.(2), and λ, the threshold in Eq.(3), on the MRR performance.

4. CONCLUSION

This paper has presented a novel best answer identification method for CQA sites. It not only uses a supporting set to bridge the lexical gap between questions and answers, but also treats them as relational data and learns an analogical reasoning-based model to rank the answers of a new question. The best answer is identified as the top-ranked one which has the strongest analogy to the positive links in the supporting set. Experiments based on 29.8 million real q-a threads showed the effectiveness of the proposed method.

5. REFERENCES

[1] Bian, J., Liu, Y., et al. Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media. In Proc. WWW, 2008.

[2] Ghahramani, Z., and Heller, K.A. Bayesian Sets. In Proc. NIPS, 2005.

[3] Jeon, J., Croft, W., et al. A Framework to Predict the Quality of Answers with Non-Textual Features. In Proc. SIGIR, 2006.

[4] Silva, R., Heller, K.A., et al. Analogical Reasoning with Relational Bayesian Sets. In Proc. AISTATS, 2007.