Pattern Recognition Letters 58 (2015) 29–34

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Predicting the quality of user-generated answers using co-training in community-based question answering portals ☆

Bingquan Liu*, Jian Feng, Ming Liu, Haifeng Hu, Xiaolong Wang

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 14 July 2014
Available online 5 March 2015

Keywords:
Co-training
Semi-supervised methods
Answer quality predicting
Surface linguistic features
Social features

ABSTRACT

Predicting the quality of user-generated answers is definitely of great importance for community-based question answering (CQA) due to the frequent occurrence of low-quality answers. Most existing answer quality prediction works combine non-textual features of user-generated answers directly without considering the diversity of non-textual features. In this paper, we propose two co-training approaches: random subspace split-based co-training (RSS-CoT) and content and social split-based co-training (CS-CoT) to predict the quality of answers by mining the relationships of non-textual features and unlabeled data in CQA. Our results demonstrate that both appropriate combination of non-textual features and unlabeled data can promote the prediction performance of answer quality.

1. Introduction

Recently, the emergence of Web services not only enriches the way people access and share knowledge, but also results in the accumulation of a large quantity of user-generated content (UGC). As a typical Web service, community-based question answering (CQA) portals, where users can post questions and resolve problems raised by other users, have accumulated massive user-generated question–answer pairs (QA pairs). For example, Baidu Knows,[1] the most popular Chinese CQA site, has received 340 million questions and billions of answer until October 2014. Many non-factoid questions can be solved by using user-generated QA pairs.

However, the distribution of answer quality in CQA varies greatly from high-quality to low-quality on account of subjectivity and knowledge limitation of users, the complexity of natural language expression and the imperfection of reputation system in CQA. High-quality answers are often question-relevant, informative, trustful and objective [1,2], like A1 in Table 1. In contrast, A2 and A3 in Table 1 are question-nonrelevant, uninformative and subjective. The frequent emergence of low-quality UGC affects the user experience of CQA portals and poses serious obstacles to data mining and knowledge reusing in CQA. Thus, the automatic quality assessment is crucial in CQA.

To predict the quality of answers is to distinguish high-quality answers and low-quality answers, so we take it as a classification task. Generally textual features are the most common features used to predict answer quality [3]. However, due to the extreme sparsity of notional word features of user-generated answers, textual features based on words are not suitable to estimate the quality of answers [4]. In contrast, non-textual features are mainly based on statistical characteristics of answer content and social information, so the problem of feature sparsity is relieved. Jeon et al. [1] and Zhou et al. [5] explored the effectiveness of non-textual features in distinguishing answer quality and got well performance. Thus, in this paper, we attempt to identify high quality answers using non-textual features.

In the task of predicting answer quality, previous studies have made some promotions, but there are still two challenges: first, these works integrate features directly without considering the relationship of features; second, a large amount of labeled data are required to train model for predicting answer quality. Generally there are other high-quality answers except best answer in CQA, but labeling answers manually is challenging. This brings the problem of lacking labeled data to train quality prediction model. For the above problems, we propose corresponding methods as follows:

(1) Taking account of the relationship between non-textual features, we propose two methods of feature space division: random subspace split (RSS); content surface linguistic features and social features split (CS), then we use ensemble learning to combine classifiers on different feature subspace to make full use of all non-textual features. The above process can find out relationship between features by feature space division and integrates all features by combining classifiers on different feature subspace.

☆ This paper has been recommended for acceptance by M. Kamel.

* Corresponding author. Tel.: +86 186 4613 9426.
E-mail address: liubq@hit.edu.cn (B. Liu).

[1] http://zhidao.baidu.com.

http://dx.doi.org/10.1016/j.patrec.2015.02.015
0167-8655/

Table 1
Example of high and low quality answers in Baidu Knows.

	Q: If we want to visit Yunnan in summer, which spots should we go to then.[a]
High-quality	A1: Generally there are two routes, Kunming, Lijiang, Xianggelila.Or Kunming, Lijiang, Banna, it depends on your time, tourism resources are very abundant in Yunnan.
Low-quality	A2: Search it through Baidu.
	A3: A lot of spots, I can help you to reserve ticket, it features in water and mountain.

[a] See http://zhidao.baidu.com/question/280484829.html.

(2) To address the problem of lacking labeled data, we introduce co-training (CoT) which can learn knowledge from unlabeled data to enhance the performance of predicting answer quality. There are a large amount of unlabeled answers in CQA, making use of those unlabeled answers by CoT is an effective way to solve the problem of lacking labeled data.

Combing the above two aspects, we propose random subspace split-based co-training (RSS-CoT) and content and social split-based co-training (CS-CoT) to predict answer quality. Experimental results show that the proposed approaches outperform the supervised methods (e.g. SVM and LR) and the semi-supervised methods (e.g. TSVM and self-training).

The remainder of this paper is organized as follows. Section 2 introduces related work. The proposed co-training approaches are described in detail in Section 3. Section 4 describes the experimental setup. Section 5 shows the experimental results. Finally we conclude this paper and provide directions for future work in Section 6.

2. Related work

Evaluating and predicting whether a candidate answer is high quality in CQA portals is a challenging task. One of the typical methods for this is exploring various features and employing machine learning techniques to address this problem. Jeon et al. [1] have proposed a framework to predict the quality of the answer incorporating non-textual features into a maximum entropy model. Lee et al. [6] and Blooma et al. [7] have performed similar research based on maximum entropy and logistic regression respectively. Meanwhile, Agichtein et al. [3] extracted more non-textual features from CQA and predict answer quality in Yahoo! Answers using a C4.5 decision tree. Subsequently, Agichtein et al. [3] and Bian et al. [2] leveraged a larger range of features including both structural and community features to identify high quality answers. To discover useful features for distinguishing high-quality answers from low-quality answers, Shah and Pomerantz [8] proposed a classification approach based on 13 different quality criteria to predict answer quality and found that answerer's information and the order of the answer in the list are the most significant features in answer quality prediction.

Predicting answer quality using the relationship between users in a community has become a widely used method. Bian et al. [2] proposed a coupled mutual reinforcement semi-supervised method to evaluate the quality of CQA content, and conducted an experiment focusing on the quality of answers in the community. Suryanto et al. [9] also employed the relationship between users, predicting the quality of answers by estimating the value of a user's authority.

Most research mentioned above is based on the mining of non-textual features. The results suggest that non-textual features are definitely of great importance to predict answer quality. By contrast, there are relatively few studies combining textual features. Wang et al. [10] conducted valuable research on deep semantic mining, proposing a deep belief network (DBN) based on a QA reconstruct and joint distribution, and they translated the discovery of high quality answers into semantic computing of QA pairs. The experimental results show that DBN can boost the performance of high-quality prediction.

This paper also uses non-textual features to identify high-quality answer, but different from the previous works, this paper proposes co-training based approaches to predict the quality of answers by mining relationships of features and knowledge of unlabeled data.

Algorithm 1: Co-training algorithm.

Input:

- Labeled dataset $\mathcal{L}$ & Unlabeled dataset $\mathcal{U}$
- Initial classifiers $\mathcal{C}_1^0$ and $\mathcal{C}_2^0$

Output: Classifiers $\mathcal{C}_1^\star$ and $\mathcal{C}_2^\star$

```
for i = 1 to K do
    Build classifier C_1^i using L in one view, predicate U .
    Pick up P positive examples and N negative examples
    with highest confidence → E^1.
    Build classifier C_2^i 2 using L in another view, predicate U .
    Pick up P positive examples and N negative examples
    with highest confidence → E^2.
    Remove E^1 ⋃ E^2 from U and add them to L .
end
C_1^⋆ = C_1^K, C_2^⋆ = C_2^K
```

3. Co-training for high-quality answers prediction

3.1. Co-training

Most classification tasks require large amounts of training examples for classifier construction. To minimize the effort of labeling examples, Blum and Mitchell [11] proposed a co-training algorithm that requires only a small number of training examples and is capable of learning from a large number of unlabeled examples. Co-training is applicable to datasets in which each data instance x can be described using two sets of features $x1$ and $x2$ (also known as two views). Two basic assumptions should be fulfilled in co-training: first, $x1$ and $x2$ are sufficient for correct classification individually, and second, $x1$ and $x2$ are conditionally independent given the category label. Theoretical analyses of the two conditions are given in Balcan et al. [12] and Blum and Mitchell [11]. The basic framework of the standard co-training algorithm is shown in Algorithm 1.

In some cases, the view-independency assumption can be relaxed within the deep research of co-training [13,14]. So far, co-training has been successfully applied to statistical parsing [15], reference resolution [16], part-of-speech tagging [17], word sense disambiguation [18] and email classification [19].

3.2. Random subspace split based co-training

The random subspace split (RSS) strategy was proposed firstly by Ho [20] and was used to construct a decision forest. The basic idea is to project data onto a random subspace of feature space. Specifically, given the training set $\mathcal{X} = \{X_i | X_i \in \mathbb{R}^n, i = 1, 2, \ldots, l\}$ and the dimension of subspace $m (m < n)$, the RSS algorithm projects an arbitrary $X_i \in \mathcal{X}$ onto the feature subspace constructed of m dimensions through sampling, denoted as X_i^m. A random feature subspace training set $\mathcal{X}^m = \{X_i^m | X_i^m \in \mathbb{R}^m, i = 1, 2, \ldots, l\}$ is constructed by all $X_i^m, i = 1, 2, \ldots, l$. Generally, we can repeat the process mentioned above K times and construct K random subspaces of the original feature spaces.

In considering the relatively small size of the feature space used in this paper, the process is shown in Algorithm 2.