Collaborative Company Profiling: Insights from an Employee’s Perspective

Hao Lin¹, Hengshu Zhu²*, Yuan Zuo¹, Chen Zhu², Junjie Wu¹, Hui Xiong³*
¹Beihang University, {linhao2014, zuoyuan, wujj}@buaa.edu.cn
²Baidu Talent Intelligence Center, {zhuhengshu, zhuchen02}@baidu.com
³Rutgers University, hxiong@rutgers.edu

Abstract

Company profiling is an analytical process to build an in-depth understanding of company’s fundamental characteristics. It serves as an effective way to gain vital information of the target company and acquire business intelligence. Traditional approaches for company profiling rely heavily on the availability of rich finance information about the company, such as finance reports and SEC filings, which may not be readily available for many private companies. However, the rapid prevalence of online employment services enables a new paradigm — to obtain the variety of company’s information from their employees’ online ratings and comments. This, in turn, raises the challenge to develop company profiles from an employee’s perspective. To this end, in this paper, we propose a method named Company Profiling based Collaborative Topic Regression (CPCTR), for learning the latent structural patterns of companies. By formulating a joint optimization framework, CPCTR has the ability in collaboratively modeling both textual (e.g., reviews) and numerical information (e.g., salaries and ratings). Indeed, with the identified patterns, including the positive/negative opinions and the latent variable that influences salary, we can effectively carry out opinion analysis and salary prediction. Extensive experiments were conducted on a real-world data set to validate the effectiveness of CPCTR. The results show that our method provides a comprehensive understanding of company characteristics and delivers a more effective prediction of salaries than other baselines.

Introduction

Recent years have witnessed the rapid development of technologies related to enterprise management, which can help organizations to keep up with the continuously changing business world. Along this line, a crucial demand is to build effective strategies for company profiling, which is an analytical process that results in an in-depth understanding of company’s fundamental characteristics, and can therefore serve as an effective way to gain vital information of the target company and acquire business intelligence. With the help of profiling, a wide range of applications could be enabled including organization risk management (Martin and Rice 2007), enterprise integration (Hollocks et al. 1997), and company benchmarking (Knuf 2000; Alling 2002; Seong Leem et al. 2008; Kerschbaum 2008; Zhu et al. 2016).

In the past decades, traditional approaches for company profiling rely heavily on the availability of the rich finance information about the company, such as finance reports and SEC filings, which may not be readily available for many private companies. Recently, with the rapid prevalence of online employment services, such as Glassdoor, Indeed, and Kanzhun, a new paradigm is enabled for obtaining the variety of company’s information from their (former) employees anonymously via the reviews, ratings and salaries of specific job positions. This, in turn, raises the question whether it is possible to develop company profiles from an employee’s perspective. For example, we can help companies to identify their advantages and disadvantages, and to predict the expected salaries of different job positions for rival companies.

However, the heterogeneous characteristic of this public information imposes significant challenges to discover typical patterns of companies during profiling. To this end, in this paper we propose a model named Company Profiling based Collaborative Topic Regression (CPCTR) to formulate a joint optimization framework for learning the latent patterns of companies, which can collaboratively model both the textual information (e.g., review) and numerical information (e.g., salary and rating). With the identified patterns, including the positive/negative opinions and the latent variable that influences salary, we can effectively carry out opinion analysis and salary predictions for different companies. Finally, we conduct extensive experiments on a real-world data set. The results show that our algorithm provides a comprehensive interpretation of company characteristics and a more effective salary prediction than other baselines. Particularly, by analyzing the results obtained by CPCTR, many meaningful patterns and interesting discoveries can be observed, such as welfare and technology are the typical pros of Baidu, while those of Tencent are training and learning.

Related Work

The related work of this paper can be grouped into two categories, namely topic modeling for opinion analysis and matrix factorization for prediction. Probabilistic topic models are capable of grouping seman-
Preliminaries

In this section, we introduce some preliminaries used throughout this paper, including data description and problem definition.

Data Description

In this paper, we aim to leverage the data collected from online employment services for company profiling. To facilitate the understanding of our data, we show a page snapshot of Indeed1 in Figure 1. Specifically, each company has a number of reviews posted by its (former) employees, each of which contains the poster’s job position (e.g., software engineer), textual information about the advantages and disadvantages of the company, and a rating score ranging from 1 to 5 to indicate the preferences of employees towards this company. Moreover, the salary range of each job position is also included for each company.

Company
Microsoft Employee Reviews for Software Engineer

Job Title
Software Engineer

Location

Rating

Job Position
Productive fun place to work with great people

Advantage

Disadvantage

Company

Salary

Company

Figure 1: An example of data description.

Problem Statement

Suppose we have a set of $E$ companies and a set of $J$ job positions. For each company $e$, there are many reviews referred to it. In each review, we have its reviewer’s role (i.e., the reviewer’s job position), rating, and two independent textual segments (positive opinion and negative opinion). Moreover, we have the average salary for each job position. For simplicity, we group reviews by their job positions and denote two words lists as $\{w_{m,j,e}\}_{m=1}^N$ and $\{w_{m,j,e}\}_{m=1}^M$ to represent positive opinion and negative opinion for a specific job position $j$, respectively.

Our problem is how to discover the latent representative patterns of job-company pair. To be more specific, there are two major tasks in this work: 1) how to learn positive and negative opinion patterns, $(\beta_j, \varphi_j)$, for each job position; and 2) how to use the latent patterns to predict job salaries ($s_{j,e}$), for each job-company pair.

Thus, we propose a model, CPCTR, for jointly modeling the numerical information (i.e., rating and salary) and review content information simultaneously. To be more specific, we use probabilistic topic model to mine review content information and use matrix factorization to handle numerical information. In terms of review content information, $\beta_j$ and $\varphi_j$ are represented by sets of opinion-related topics. Besides, each job-company pair $(j, e)$ has a topic pattern $\theta_{j,e}$, indicating its probability over $\beta_j$ and $\varphi_j$. In terms of numerical information, we use a low-dimensional representation derived from numerical information, such as salary and rating, to represent job position and combine it with $\theta_{j,e}$ to model the latent relationship among them.

Obviously, our model is a combination of probabilistic topic modeling and matrix factorization, similar to CTR. However, unlike CTR that only learns a global topic-word distribution $\beta$ and topic proportion $\theta_j$ for each item $j$, our model can learn two kinds of job related topic-word patterns, including a positive topic-word distribution $\beta_j$ and a negative topic-word distribution $\varphi_j$. Moreover, in contrast with CTR, which cannot incorporate both rating and salary information into one optimization model simultaneously, our method can model these two numerical values and utilize the learned opinion patterns for more precise salary prediction.
Thus, our model leads to a more comprehensive interpretation of company profiling and provides a collaborative view from opinion modeling to salary prediction.

**Technical Details**

In this section, we formally introduce the technical details of our model CPCTR.

**Model Formulation**

As mentioned above, our model CPCTR is a Bayesian model which combines topic modeling with matrix factorization. The graphical representation of CPCTR is shown in Figure 2. To facilitate understanding, we look into the model in two sides.

On one side, we model the job-company pair with a latent topic vector $\theta_{j,e} \in R^K$, where $K$ is the number of topics. In probabilistic topic modeling, job position $j$ can be represented by two latent matrices, i.e., the positive opinion topics $\beta_j \in R^{K \times G}$ and the negative opinion topics $\varphi_j \in R^{K \times G}$, where $G$ is the size of vocabulary. For the $n$th word $w_{n,j,e}$ in a positive review of job-company pair, we assume there is a latent variable denoted as $z_{n,j,e} \in \{1, \ldots, K\}$, indicating the word’s corresponding topic. To be more specific, given $z_{n,j,e} = k$, $w_{n,j,e} \in \theta_{k,j}^T$. Meanwhile, the positive latent pattern $z_{n,j,e}$ is considered to be drawn from the multinomial distribution $Mult(\theta_{j,e})$. A similar process can be conducted for the negative review.

On the other side, we conduct matrix factorization for salary prediction. In matrix factorization, we represent job position and job-company pair in a shared latent low-dimensional space of dimension $K$, i.e., job position $j$ is represented by latent vectors $u_j \in R^K$ and $b_j \in R^K$, which indicate the influences of job positions over salary and rating, respectively. Similarly, the job-company pair $(j, e)$ is represented by a latent vector $v_{j,e} \in R^K$, which indicates the joint influences of job-company pair over numeric rating and salary values. Here, we assume the latent vector $u_j$ and $b_j$ follow Gaussian distributions with parameters $(0, \lambda_u^{-1} I_K)$ and $(0, \lambda_b^{-1} I_K)$, respectively. And, the latent vector $v_{j,e}$ is derived from $\theta_{j,e}$ by adding an offset, $\epsilon$, which follows a Gaussian distribution with parameters $(0, \lambda_v^{-1} I_K)$. Therefore, it is obvious that $v_{j,e}$ is the key point by which we jointly model both content and numerical information.

We form the prediction of salary values of a specific job-company pair $(j, e)$ through the inner product between their latent representations, i.e.,

$$\hat{s}_{j,e} = u_j^T v_{j,e}.$$  \(1\)

Note that in our model, we first group reviews, ratings, and salary information by the posters’ job-company pair. We then calculate the average ratings, average salaries and aggregate reviews as one single document for each job-company pair. The complete generative process of our model is demonstrated in Algorithm 1. In the following, we leverage the Bayesian approach for parameter learning.

**Parameter Learning**

In the above generative process, we denote mathematical notations as follows. $S = \{s_{j,e} \}_{j=1,e=1}^{J,E}$, $R = \{r_{j,e} \}_{j=1,e=1}^{J,E}$, $W_P = \{w_{n,j,e} \}_{n=1,j=1,e=1}^{N,J,E}$, $W_C = \{w_{m,j,e} \}_{m=1,j=1,e=1}^{M,J,E}$, $\theta = \{\theta_{j,e} \}_{j=1,e=1}^{J,E}$, $U = \{u_j \}_{j=1}^{J}$, $B = \{b_j \}_{j=1}^{J}$, $V = \{v_{j,e} \}_{j=1,e=1}^{J,E}$. The joint likelihood of data, i.e., $S, R, W_P, W_C$, and the latent factors $\theta, \beta, \varphi, U, B, V$ under the full
model is
\[ p(S, R, W^P, W^C, \theta, \beta, \phi, U, B, V | \lambda_\star, \alpha, \tau, \rho) \]

\[ = p(U | \lambda_u)p(B | \lambda_b)p(V | \lambda_v, \theta)p(R, S | U, B, V, \lambda_v, \lambda_u) \]

\[ p(W^P, W^C | \beta, \phi, \theta)p(\theta | \alpha)p(\beta | \tau)p(\phi | \rho) \]

\[ = \prod_j p(u_j | \lambda_u) \prod_j p(b_j | \lambda_b) \prod_{e,j} p(v_{j,e} | \lambda_v, \theta, \beta, \phi) \]

\[ \prod_{e,j} p(r_{j,e} | b_j, v_{j,e}, \lambda_r) \prod_{e,j} p(s_{j,e} | u_j, v_{j,e}, \lambda_s) \]

\[ \prod_{e,j} \left( \prod_n p(w_{n,j,e} | \beta, \phi, \theta) \prod_m p(w_{m,j,e}^C | \phi, \theta) \right) \]

\[ p(\theta | \alpha)p(\beta | \tau)p(\phi | \rho). \tag{2} \]

For learning the parameters, we develop an EM-style algorithm to learn the maximum a posterior (MAP) estimation. Maximization of posterior is equivalent to maximizing the complete log likelihood of \( R, S, W^P, W^C, \theta, U, B, V \), given \( \lambda_\star, \beta, \phi \),

\[ L = -\frac{\lambda_0}{2} \sum_j b_j^T b_j - \frac{\lambda_u}{2} \sum_j u_j^T u_j - \frac{\lambda_s}{2} \sum_{j,e} (r_{j,e} - b_j v_{j,e})^2 \]

\[ - \frac{\lambda_r}{2} \sum_j (s_{j,e} - u_j v_{j,e})^2 - \frac{\lambda_s}{2} \sum_{j,e} (v_{j,e} - \theta_j)^2 \]

\[ + \sum_{j,e} \log(p(v_{j,e} | \beta, \phi, \theta)) \sum_n \log(p(w_{n,j,e} | \beta, \phi, \theta)) \]

\[ + \sum_m \log(p(w_{m,j,e}^C | \phi, \theta)) \sum_k \log(p(\theta_{k,j,e} | \phi, \beta, \theta)). \tag{3} \]

Here, we employ coordinate ascent (CA) approach to alternately optimize the latent factors \( \{u_j, b_j, v_{j,e}\} \) and the simplex variables \( \theta_{j,e} \) as topic proportion. For \( u_j, b_j, v_{j,e} \), we follow in a similar fashion as for basic matrix factorization (Hu, Koren, and Volinsky 2008). Given the current estimation of \( \theta_{j,e} \), taking the gradient of \( L \) with respect to \( u_j, b_j, v_{j,e} \) and setting it to zero leads to

\[ u_j = (\lambda_0 I_K + \lambda_s \sum_{j,e} v_{j,e} v_{j,e}^T)^{-1} \lambda_s \sum_{j,e} v_{j,e} s_{j,e} \tag{4} \]

\[ b_j = (\lambda_0 I_K + \lambda_r \sum_{j,e} v_{j,e} v_{j,e}^T)^{-1} \lambda_r \sum_{j,e} v_{j,e} r_{j,e} \tag{5} \]

\[ v_{j,e} = (\lambda_0 I_K + \lambda_s u_j u_j^T + \lambda_r b_j b_j^T)^{-1} \]

\[ (\lambda_0 \theta_{j,e} + \lambda_s u_j s_{j,e} + \lambda_r b_j r_{j,e}). \tag{6} \]

Given \( U, B, V \), we then apply a variational EM algorithm described in LDA (Blei, Ng, and Jordan 2003) to learn the topic proportion \( \theta_{j,e} \). We first define \( q(z_{n,j,e} = k) = \phi^P_{k,n,j,e} \) and \( q(z_{m,j,e} = k) = \phi^C_{k,m,j,e} \) and then separate the items that contain \( b_j \) and apply Jensen’s inequality,

\[ \mathcal{L}(\theta_{j,e}) = -\frac{\lambda_0}{2} (v_{j,e} - \theta_{j,e})^T (v_{j,e} - \theta_{j,e}) + \]

\[ \sum_n \phi^P_{k,n,j,e} \log q(\theta_{j,e} | \beta_{k,n,j,e}) - \log \phi^P_{k,n,j,e} + \]

\[ \sum_m \phi^C_{k,m,j,e} \log q(\phi_{k,m,j,e} | \phi_{k,m,j,e}) - \log \phi^C_{k,m,j,e} \]

\[ \mathcal{L}(\theta_{j,e}, \phi_{j,e}). \tag{7} \]

where \( \phi_{j,e} = \{\phi^P_{k,n,j,e} k=1,n=1 \} \). In the E-step, the optimal variational multinomial \( \phi^P_{k,n,j,e} \) and \( \phi^C_{k,m,j,e} \) satisfy

\[ \phi^P_{k,n,j,e} \propto \theta_{k,n,j,e} \beta_{k,n,j,e} \tag{8} \]

\[ \phi^C_{k,m,j,e} \propto \theta_{k,m,j,e} \phi_{k,m,j,e} \tag{9} \]

The \( \mathcal{L}(\theta_{j,e}, \phi_{j,e}) \) gives a tight lower bound of \( \mathcal{L}(\theta_{j,e}) \). Similar to CTR (Wang and Blei 2011), we use projection gradient [Bertsekas 1999] to optimize \( \beta_{j,e} \). Coordinate ascent can be applied to optimize remaining parameters \( U, B, V, \theta \) and \( \phi \). Then following the same M-step for topics in LDA [Blei, Ng, and Jordan 2003], we optimize \( \beta \) and \( \phi \) as follows,

\[ \beta_{g,k,j} \propto \sum_n \phi^P_{k,n,j,e} \mathbb{1}[w_{n,j,e} = g] \tag{10} \]

\[ \phi_{g,k,j} \propto \sum_n \phi^C_{k,m,j,e} \mathbb{1}[w_{m,j,e} = g], \tag{11} \]

where we denote \( g \) as an arbitrary term in the vocabulary set.

**Discussion on Salary Prediction**

After all the optimal parameters are learned, the CPCTR model can be used for salary prediction by Equation [1]. In this task, rating values and review content of the predicted job-company pair \( (j, e) \) are available, but no salary information of \( (j, e) \) pair is available. To obtain the topic proportion \( \theta_{j,e} \) for the predicted job-company pair \( (j, e) \), we optimize Equation [7].

In particular, we only focus on the task of salary prediction, although rating prediction can be conducted in a similar way. Since reviews are always accompanied by ratings, ratings should be regarded as part of opinion information. Therefore, in this work we treat the ratings as the complementary of reviews for opinion mining, and the side information for salary prediction.

**Experimental Results**

In this section, we first give a short parameter sensitivity discussion to show the robustness of our model and then evaluate the salary prediction performance of CPCTR based on a real-world data set with several state-of-the-art baselines. Finally, we empirically study the pros and cons for each job-company pair learned from their employees’ review.

**Experimental Setup**

**Data Sets.** Kanzhun\(^3\) is one the largest online employment website in China, where members can review companies and assign numeric ratings from 1 to 5, and post their own salary information. Thus, Kanzhun provides an ideal data source for experiments on company profiling and salary prediction. The data set used in our experiments consists of 934 unique companies which contains at least one of total 1,128 unique job positions, i.e., for a specific company, at least one job’s average salary and rating has been included. Moreover, the

\(^3\)http://www.kanzhun.com/
Table 1: The prediction performance of different methods under 5-fold cross-validation.

| Fold | Method | CPCTR | PMF | RSVD | CTR |
|------|--------|-------|-----|------|-----|
|      | rMSE   | 0.0528| 0.0561| 0.0608| 0.0670|
| 1    | MAE    | 0.0347| 0.0356| 0.0419| 0.0433|
|      | rMSE   | 0.0530| 0.0518| 0.0592| 0.0597|
| 2    | MAE    | 0.0346| 0.0332| 0.0401| 0.0394|
|      | rMSE   | 0.0506| 0.0499| 0.0595| 0.0621|
| 3    | MAE    | 0.0328| 0.0322| 0.0413| 0.0414|
|      | rMSE   | 0.0680| 0.0703| 0.0743| 0.0815|
| 4    | MAE    | 0.0345| 0.0365| 0.0425| 0.0472|
|      | rMSE   | 0.0479| 0.0514| 0.0543| 0.0609|
| 5    | MAE    | 0.0332| 0.0354| 0.0407| 0.0433|
|      | Average rMSE | 0.0545| 0.0559| 0.0616| 0.0662|
|      | Average MAE | 0.0340| 0.0346| 0.0413| 0.0429|

Baseline Methods. To evaluate the performance of salary prediction for CPCTR, we chose three state-of-the-art benchmark methods for comparisons, including PMF (Salakhutdinov and Mnih 2007), Regularized Singular Value Decomposition of data with missing values RSVD¹ and Collaborative Topic Regression CTR (Wang and Blei 2011).

Evaluation Metrics. We used two widely-used metrics, i.e., root Mean Square Error (rMSE), Mean Absolute Error (MAE), for measuring the prediction performance of different models. Specifically, we have

\[
RMSE = \sqrt{\frac{1}{N} \sum (s_i - \hat{s}_i)^2}
\]

(12)

\[
MAE = \frac{1}{N} \sum |s_i - \hat{s}_i|
\]

(13)

where \(s_i\) is the actual salary of \(i\)th job-company pair, \(\hat{s}_i\) is its salary prediction and \(N\) is the number of test instances.

Experimental Settings. In our experiments, we used 5-fold cross-validation. For every job position that was posted by at least 5 companies, we evenly split their job-company pairs (average rating/salary values) into 5 folds. We iteratively considered each fold to be a test set and the others to be the training set. For those job positions that were posted by fewer than 5 companies, we always put them into the training set. This leads to that all job positions in the test set must have appeared in the training set, thus it guarantees the in-matrix scenario for CTR model in prediction. For each fold, we fitted a model to the training set and test on the within-fold jobs for each company. Note that, each company has a different set of within-fold jobs. Finally, we obtained the predicted salaries and evaluated them on the test set.

The parameter settings of different methods are stated as follows. For all methods, we set the number of latent factor to \(K = 5\) and the maximum iterations for convergence as \(maxIter = 500\). For probabilistic topic modeling in CTR and CPCTR, we set the parameters \(\alpha_{smooth} = 0.01\). For CTR, we used fivefold cross validation to find that \(\lambda_u = 10\), \(\lambda_v = 0.01\), \(a = 1\) and \(b = 0.01\) provides the best performance. For our model CPCTR, we chose the parameters by using grid search on held out predictions. As a default setting for CPCTR, we set \(\lambda_u = 0.1\), \(\lambda_b = 0.01\), \(\lambda_v = 1000\), \(\lambda_r = 1\), \(\lambda_s = 1\). More detailed discussions about parameter sensitivity of our model will be given in the following subsection. Additionally, for convenience of parameter choosing, we used min-max method to normalize all values of rating/salary into \([0, 1]\) range.

Parameter Sensitivity

In our model, the content parameter \(\lambda_v\) controls the contribution of review content information to model performance and the rating parameter \(\lambda_r\) balances the contribution of rating information to model performance. In the left plot of

Figure 3: Plots of prediction performance for CPCTR by varying content parameter \(\lambda_v\) and rating parameter \(\lambda_r\).

Figure 4: Topics of job position “software engineer”.

¹https://github.com/alabid/PySVD
In this paper, we proposed a model CPCTR for company profiling, which can collaboratively model the textual information and numerical information of companies. A unique perspective of CPCTR is that it formulates a joint optimization framework for learning the latent patterns of companies, including the positive/negative opinions of companies and the latent topic variable that influences salary from an employee’s perspective. With the identified patterns, both opinion analysis and salary prediction can be conducted effectively. Finally, we conducted extensive experiments on a real-world data set. The results showed that our model provides a comprehensive interpretation of company characteristics and a more effective salary prediction than baselines.

Concluding Remarks

Here, we apply CPCTR to carry out opinion analysis for different companies based on employees’ reviews. The objective is to effectively reveal the pros and cons of companies, which indeed helps for competitor benchmarking.

To illustrate the effectiveness of learning job-position level topic-word distributions, we listed 3 positive topics and 3 negative topics of job position Software Engineer inferred from CPCTR, as shown in Figure 5. Each topic is represented by 5 most probable words for that topic. It can be seen that our method has an effective interpretation of latent job position pattern and these topics accurately capture the common semantics of the job position Software Engineer in the whole market. We can see some interesting positive/negative topic patterns. For positive topics, topic 0 is about job environment, topic 1 is about flexible work time, topic 2 is technology atmosphere. For negative topics, topic 0 is about overtime, topic 1 is about prospect and promotion, and topic 2 is about opportunity and welfare.

We also compared the pros and cons among BAT, which is the abbreviation of three largest and most representative Chinese Internet companies, i.e., Baidu, Alibaba and Tencent. Specifically, we presented the pros and cons with most probable words appearing in learned topics for each company, given the job position Software Engineer in Figure 5. As can be seen, topics for each job-company pair can effectively capture the specific characteristics of each company. For instance, the typical pros of Baidu are welfare and technology, while those of Tencent are training and learning and those of Alibaba are culture and atmosphere. Interestingly, employees of all these three companies chose overtime as their cons, and the management of Tencent seems to be a typical cons.

Empirical Study of Opinion Profiling

Figure 5: Pros and cons of various enterprises given job position “software engineer”.

Figure 3 we vary the content parameter $\lambda_r$ and rating parameter $\lambda_v$ from $1 \times 10^{-5}$ to $1 \times 10^5$ to study the effect on the performance of salary prediction, and the average performance within fivefold cross validation is shown in this plot. First, we can see that CPCTR shows good prediction performance when $\lambda_r >= 1$ and $\lambda_v >= 1$, and achieves the best prediction accuracy when $\lambda_r = 1000$ and $\lambda_v = 1$, which is the default setting for CPCTR. Next, for facilitating comparison, we shrink the range of $\lambda_r$ and $\lambda_v$ into $[1, 1e+05]$ and show the right plot of Figure 3. From this plot, we can see that almost all cases of CPCTR outperform other state-of-the-art baselines, except for $\lambda_v = 1$. The results show small and negligible fluctuation with varied $\lambda_v$ and $\lambda_r$, and CPCTR becomes insensitive to these two parameters.

Performance of Salary Prediction

We show the prediction performance of different methods in Table 1. Note that, the best results are highlighted in bold and the runner-up are denoted in italic. From the results, we could observe that CPCTR achieves the best average prediction performance in terms of 5-fold cross-validation, and outperforms other baselines in three folds. This is in great contrast to CTR, which shows poor prediction performance in all five folds. It is because that, although CTR can integrate textual information for salary prediction, it cannot utilize the rating information and does not explicitly model the positive/negative topic-word distribution. Among traditional collaborative filtering methods, PMF consistently outperforms RSV in all five folds, which demonstrates the effectiveness of probabilistic methods on prediction tasks. Based on the above analysis, CPCTR can be regarded as a more comprehensive and effective framework for company profiling, that can integrate review opinions and ratings for salary prediction.

Acknowledgments

This work was partially supported by NSFC (71322104, 71531001, 71471009, 71490723), National High Technology Research and Development Program of China (SS2014AA012303), National Center for International Joint Research on E-Business Information Processing (2013B01035), and Fundamental Research Funds for the Central Universities.
